Konvertierung bibliografischer Referenzdaten in ein neutrales Austauschformat: Probleme und Lösungsmöglichkeiten am Beispiel der Datenbank "Literatur zur Informationserschließung"

Bachelorarbeit zur Erlangung des akademischen Grades
Bachelor of Arts
im Studiengang Bibliothek und digitale Kommunikation
an der Fakultät für Informations- und Kommunikationswissenschaften
der Technischen Hochschule Köln

vorgelegt von: Kevin Blischke

eingereicht bei: Prof. Dr. Klaus Lepsky
Zweitgutachterin: Prof. Dr. Mirjam Blümm

Köln, 11.03.2024

Abstract

Die Konvertierung von bibliographischen Daten in andere Formate stellt eine häufige Herausforderung in der bibliothekarischen Arbeit dar, wie die Systemumstellung vieler Bibliotheken auf das Bibliotheksmanagementsystem Alma zeigt. Dabei ist die verlustfreie Durchführung dieses Prozesses eine besondere Schwierigkeit, die aus der Verschiedenheit der Formate resultiert. Ein konkretes Beispiel für eine solche zu konvertierende Datenmenge ist die Literaturdatenbank "Literatur zur Informationserschließung", welche 44.218 bibliographische Einträge enthält und von einer modifizierten Form des Allegro-Neutralformats in das RIS-Format konvertiert werden soll. Dabei wird auf der Grundlage von erarbeiteten Konkordanzen zwischen beiden Formaten und Untersuchungen der Datenbank mit regulären Ausdrücken, sowie einem Pythonskript ein Programm geschrieben, das die Datenbank in das Zielformat konvertieren soll. Das Ergebnis wird anhand einer proportionalen Schichtenstichprobe evaluiert. Abschließend werden der Entwicklungsprozess und das Ergebnis hinsichtlich des stattgefundenen Informationsverlustes bei dem Konvertierungsprozess reflektiert.

Schlagwörter: Bibliographische Datenkonvertierung, Datenkonvertierungsprogramm, Allegro-Neutralformat, RIS-Format

The conversion of bibliographic data into other formats is a frequent challenge in library work, as the system conversion of many libraries to the Alma library management system shows. The loss-free implementation of this process is a particular difficulty resulting from the diversity of formats. A concrete example of such a data set to be converted is the literature database "Literature on information indexing", which contains 44,218 bibliographic entries and is to be converted from a modified form of the Allegro neutral format into the RIS format. Based on concordances between the two formats and investigations of the database with regular expressions and a Python script, a program is written to convert the database into the target format. The result is evaluated using a proportional stratified random sample. Finally, the development process and the result are reflected with regard to the loss of information during the conversion process.

Keywords: Bibliographic data conversion, Data conversion program, Allegro neutral format, RIS format

Inhaltsverzeichnis

Abstract ..... **I**
Tabellenverzeichnis ..... **III**
Abbildungsverzeichnis ..... **IV**
1 Einleitung **1**
1.1 Zielsetzung 1
1.2 Vorgehen 2
1.3 Methodik 3
2 Beschreibung der Formate **5**
2.1 Modifiziertes Allegro-Neutralformat 5
2.2 RIS-Format 10
3 Konkordanz zwischen den Formaten **16**
3.1 Dokumenttypen 16
3.2 Felder 18
4 Untersuchung der Datenbank **22**
5 Programm zur Konvertierung der Datenbank **27**
5.1 Beschreibung 27
5.2 Implementierung 33
6 Ergebnis der Konvertierung **35**
7 Fazit **39**
Literaturverzeichnis **41**
Erklärung **44**

Tabellenverzeichnis

Tabelle 1: Felder des modifizierten Allegro-Neutralformats 8

Tabelle 2: Dokumenttypen des modifizierten Allegro-Neutralformats 9

Tabelle 3: Felder des RIS-Formats 12

Tabelle 4: Dokumenttypen des RIS-Formats 14

Tabelle 5: Konkordanz zwischen Dokumenttypen 17

Tabelle 6: Konkordanz zwischen Feldern 21

Tabelle 7: Anteile der Dokumenttypen in der Datenbank "Literatur zur Informationserschließung" 36

Tabelle 8: Ergebnisse der Stichprobenprüfung des Konvertierungsprozesses 37

Abbildungsverzeichnis

Abbildung 1: Aktivitätsdiagramm des Konvertierungsprogramms 29

1 Einleitung

Die Konvertierung von bibliographischen Daten in andere Formate stellt eine häufige Herausforderung in der bibliothekarischen Arbeit dar, wie die Systemumstellung vieler Bibliotheken auf das Bibliotheksmanagamentsystem Alma zeigt.[1] Dabei ist die verlustfreie Durchführung dieses Prozesses eine besondere Schwierigkeit, die aus der Verschiedenheit der Formate resultiert.[2]

Bei dieser Problematik existiert keine allgemeingültige Musterlösung, jede Datenmenge muss individuell betrachtet werden. Daher sind maßgeschneiderte Lösungen erforderlich, wenn nicht mit Standardsystemen gearbeitet wird. Bei diesen kann das Vorhandensein von entsprechenden Tools ansonsten angenommen werden, wie beispielsweise dem Datenaustauschformat MARC 21.[3] Die Herangehensweise wird hierbei von dem Ausgangsformat der Daten, das Zielformat, die spezifischen Anforderungen an die Konvertierung, sowie die bisherige Handhabung der Datenmenge bestimmt.

Ein konkretes Beispiel für eine zu konvertierende Datenmenge ist die Literaturdatenbank "Literatur zur Informationserschließung".[4] Diese liegt als DBM-Datei vor,[5] welche eine Midos-Datenbankdatei ist und als Text-Datei genutzt werden kann. Die Datenbank enthält derzeit 44.218 bibliographische Einträge von Literatur über Informationserschließung,[6] welche in einer modifizierten Form des Allegro-Neutralformats gespeichert sind.[7] Diese Datenbank soll nun in das RIS-Format konvertiert werden.

Bisher existiert kein spezialisiertes Programm zur Konvertierung des Allegro-Neutralformats in das RIS-Format. Daher erfordert diese Aufgabe eine sorgfältige Analyse und die Entwicklung einer maßgeschneiderten Konvertierungslösung.

1.1 Zielsetzung

Wie bereits dargelegt hat die vorliegende Bachelorarbeit das Ziel, die Datenbank "Literatur zur Informationserschließung" von einer modifizierten Version des Allegro-

---

[1] Vgl. Plaum, C. (08.11.2022). *Die Hochschulbibliotheken in NRW auf dem Weg in die Alma-Cloud.* ABI Technik, 42(4), 265–271. https://doi.org/10.1515/abitech-2022-0046. S.267ff.; Kann, B. (24.11.2018). Alma im Österreichischen Bibliothekenverbund (OBV): Aus der Werkstatt der OBVSG. Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare, 71(2), 307–319. https://doi.org/10.31263/voebm.v71i2.2133. S.309ff.

[2] Vgl. Beiler, C., Gratzl, P., Schubert, B., Steiner, C., & Steltzer, R. (24.11.2018). *Erschließungsarbeit in Alma: Erfahrungen aus dem OBV vor, während und nach der Aleph-Ablöse.* Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare, 71(2), 282–306. https://doi.org/10.31263/voebm.v71i2.2134. S.285f.

[3] Vgl. Library of Congress Network Development and MARC Standards Office. (o.D.). MARC Specialized Tools. MARC Records, Systems, and Tools (Network Development and MARC Standards Office, Library of Congress). https://www.loc.gov/marc/marctools.html. [Letzter Aufruf: 11.03.2024].

[4] Vgl. Lepsky, K. (o.D.). *Literatur: Literaturdatenbank zu den Themen Informationserschließung und Information Retrieval. Indexierung-Retrieval.* https://www.indexierung-retrieval.de/2013/02/literatur.html. [Letzter Zugriff: 11.03.2024].

[5] Vgl. "litie34.dbm" in den beigefügten Anlagen.

[6] Vgl. "litie34.dbm" in den beigefügten Anlagen.

[7] Vgl. Lepsky, K. (05.2017). *Kategorienschema der Datenbank "Literatur zur Informationser-schließung".* https://ixtrieve.fh-koeln.de/lehre/kategorienschema-litac.pdf. [Letzter Aufruf: 11.03.2024]. ;

Neutralformats in das RIS-Format zu konvertieren. Diese Konvertierung wird angestrebt, um die Interoperabilität der Datenbank mit verschiedenen Systemen und Schnittstellen zu ermöglichen, sowie langfristig zur Erhaltung der Datenbank beizutragen.

Angesichts der Anzahl von 44.218 bibliographischen Einträgen ist eine manuelle Konvertierung nicht praktikabel, daher wird nach einer automatisierten Lösung gesucht. Die Entwicklung einer solchen Lösung erfordert detaillierte Kenntnisse über die Zusammensetzung und Muster innerhalb der Datenbank.

Im Verlauf dieses Konvertierungsprozesses werden die angewendeten Methoden, Herausforderungen und mögliche Lösungsansätze ausführlich dargestellt. Auf Basis dieser Zielstellung wird die zentrale Forschungsfrage formuliert: "Wie und unter welchem Informationsverlust können bibliographische Referenzdaten in ein neutrales Austauschformat konvertiert werden?". Diese Fragestellung bildet den Kern der Untersuchung und wird im weiteren Verlauf der Arbeit systematisch bearbeitet und beantwortet.

1.2 Vorgehen

Um die genannte Zielstellung zu erreichen wurde ein systematisches Vorgehen entwickelt, das in diesem Abschnitt beschrieben wird, welches im Rahmen dieser Arbeit umgesetzt werden soll.

Um ein umfassendes Verständnis für die beiden Formate, das modifizierte Allegro-Neutralformat und das RIS-Format zu erlangen, wird zunächst ihre Definition und Analyse vorgenommen. Dieser Prozess konzentriert sich hauptsächlich auf die Dokumentation der Formate. Zusätzlich werden Teile der Datenbank "Literatur zur Informationserschließung" untersucht. Dies ermöglicht Rückschlüsse auf das modifizierte Allegro-Neutralformat zu ziehen. Hierdurch soll zunächst eine definitorische Annäherung für beide Formate geschaffen werden.

Im nächsten Schritt erfolgt die Erarbeitung von Konkordanzen zwischen dem modifizierten Allegro-Neutralformat und dem RIS-Format. Dieser Prozess zielt darauf ab, herauszufinden, wie eine Übersetzung vom modifizierten Allegro-Neutralformat in das RIS-Format allgemein möglich ist. Dabei werden die strukturellen Gemeinsamkeiten und Unterschiede der beiden Formate analysiert, um einen effizienten und präzisen Übersetzungsprozess zu ermöglichen.

Des Weiteren werden im Rahmen der Untersuchung der Datenbank die Inhalte dieser genauer betrachtet. Ziel dieses Schrittes ist es herauszufinden, wie die Übersetzung der Datenbank anhand der zuvor erarbeiteten Konkordanzen konkret realisiert werden kann. Durch diese detaillierte Untersuchung werden mögliche Herausforderungen und Lösungsansätze identifiziert, um einen reibungslosen Übersetzungsprozess zu gewährleisten.

Basierend auf den erarbeiteten Konkordanzen und den Ergebnissen der Untersuchung der Datenbank wird ein Programm zur Konvertierung der Datenbank vom modifizierten

Allegro-Neutralformat in das RIS-Format entwickelt. Dieses Programm wird sprachunabhängig beschrieben, um eine breite Anwendbarkeit zu gewährleisten. Die Implementierung erfolgt in der Programmiersprache Java.[8] Das Ziel dieser Entwicklung ist es, die Übersetzung der Datenbank effektiv und zuverlässig zu realisieren, wodurch eine nahtlose Integration in verschiedene Umgebungen ermöglicht wird.

Um die Qualität des Konvertierungsprozesses zu ermitteln, wird eine stichprobenartige Kontrolle der Datensätze in der konvertierten Datenbank durchgeführt. Durch diesen Validierungsprozess wird sichergestellt, dass die Übersetzung der Datenbank fehlerfrei und präzise erfolgt ist. Etwaige Unstimmigkeiten oder Fehler können identifiziert und korrigiert werden, um eine hochwertige und verlässliche Übersetzung zu gewährleisten.

1.3 Methodik

Im Folgenden werden die Methoden dargestellt, die angewendet werden, um das zuvor beschriebene Vorgehen zu realisieren.

Es werden reguläre Ausdrücke verwendet, um Inhalte der Datenbank "Literatur zur Informationserschließung" zu untersuchen. Innerhalb dieser Arbeit wird dafür das Textverarbeitungsprogramm Notepad++ verwendet, das unter der GNU-Lizenz frei verfügbar ist.[9] Durch die Tastenkombination Strg + F kann in diesem ein Suchmenü geöffnet werden, das die Verwendung von regulären Ausdrücken ermöglicht.[10]

Bei regulären Ausdrücken handelt sich um Sequenzen von Zeichen, die ein Suchmuster bilden. Diese Muster werden genutzt, um Texte gezielt nach bestimmten Kriterien zu durchsuchen oder zu manipulieren. Das bedeutet, dass nicht nur nach exakten Übereinstimmungen gesucht werden kann, sondern auch nach Mustern und Varianten.[11] Dank dieser Funktionalität ist es möglich, die gesamte Datenbankdatei effizient nach bestimmten Zeichensequenzen und Regelmäßigkeiten zu durchsuchen und somit zu analysieren.

Für die Untersuchung von Zusammenhängen zwischen verschiedenen Inhalten der Datenbank wird ein selbstgeschriebenes Skript in der Skriptsprache Python verwendet.[12] Dieses Skript bietet die Möglichkeit, beliebig viele reguläre Ausdrücke als Eingabe zu akzeptieren, nach denen Einträge der Datenbank durchsucht werden sollen. Es analysiert dann die Datenbank und gibt die Anzahl aller Einträge zurück, in denen alle

[8] Vgl. Oracle. (o.D:). *What is Java and why do I need it?.* Java.com. https://www.java.com/en/download/help/whatis_java.html. [Letzter Aufruf: 11.03.2024].

[9] Vgl. Ho, D. (o.D.). *What is notepad++.* Notepad++. https://notepad-plus-plus.org/. [Letzter Aufruf: 11.03.2024].

[10] Vgl. Notepad++. (o.D.). *Searching: Notepad++ User Manual.* Notepad++ User Manual. https://npp-user-manual.org/docs/searching/. [Letzter Aufruf: 11.03.2024].

[11] Vgl. Microsoft. (12.06.2023). *Sprachelemente für reguläre Ausdrücke: Kurzübersicht - .NET.* Microsoft Learn. https://learn.microsoft.com/de-de/dotnet/standard/base-types/regular-expression-language-quick-reference. [Letzter Aufruf: 11.03.2024].

[12] Vgl. Python. (o.D.). *Welcome to Python.org.* https://www.python.org/about/. [Letzter Aufruf: 11.030.2024].

eingegebenen regulären Ausdrücke gleichzeitig gefunden wurden. Das Skript kann den beigefügten Anlagen in dem Ordner "util" als "util.py" entnommen werden.

Um das Skript auszuführen, muss der folgende Befehl in der Kommandozeile eingegeben werden: "python <Pfad/>util.py". Dabei muss eine Python-Laufzeitumgebung mit der Version 3.11 oder höher auf dem System vorhanden sein, um die Ausführung zu ermöglichen.[13] Es ist wichtig zu beachten, dass "<Pfad>" durch den aktuellen Dateipfad zum Programm ersetzt werden muss.

Nachdem das Skript gestartet wurde, erscheint in der Kommandozeile die Aufforderung "Field to search:". Hier kann das zu untersuchende Feld eingegeben und mit der Eingabetaste bestätigt werden. Anschließend kann hinter dem Satz "Regex in field to search:" ein regulärer Ausdruck eingegeben werden, um innerhalb des zuvor angegebenen Feldes nach bestimmten Mustern zu suchen. Eine leere Eingabe findet alle Vorkommen des Feldes. Diese Abfragen können beliebig oft wiederholt werden, bis die Abfrage "Field to search:" leer bleibt. Die Bedeutung dieser Suchabfragen für die Untersuchung der Datenbank wird in dem Kapitel "4 Untersuchung der Datenbank" genauer erläutert.

In demselben Ordner, in dem sich das Skript befindet, ist zusätzlich ein Test mit dem Namen "test_util.py", der die Funktionalität des Skripts sicherstellt. Dieser ist ein Komponententest, welcher Teile des Skripts isoliert und sie unter verschiedenen Testszenarien ausführt.[14] Zur Ausführung des Tests muss folgender Befehl in der Kommandozeile verwendet werden: "python <Pfad/>test_util.py". Dabei muss "<Pfad\>" durch den aktuellen Dateipfad zum Programm ersetzt werden. Wenn keine Fehlermeldung angezeigt wird, war der Test erfolgreich.

Zusätzlich werden Einträge der Datenbank intellektuell untersucht, wenn eine algorithmische Untersuchung nicht möglich ist. Um eine Bewältigung der Datenmenge zu ermöglichen wird in solchen Fällen mit Stichproben gearbeitet.

[13] Vgl. Python. (10.2023). *Python release python 3.11.6.* https://www.python.org/downloads/release/python-3116/. [Letzter Aufruf: 11.03.2024].

[14] Vgl. Microsoft. (29.11.2024). *Durchführen von Komponententests Mithilfe des test-explorers: visual studio (windows).* Microsoft Learn. https://learn.microsoft.com/de-de/visualstudio/test/unit-test-basics?view=vs-2022. [Letzter Aufruf: 11.030.2024].

2 Beschreibung der Formate

Zunächst erfolgt eine detaillierte Beschreibung, sowohl des modifizierten Allegro-Neutralformats, als auch des RIS-Formats. Diese Beschreibungen dienen als grundlegende Referenzpunkte für die Erstellung von Konkordanzen zwischen den beiden Formaten, die noch erarbeitet werden sollen. Durch die genaue Erfassung der Struktur und der Merkmale jedes Formats wird eine Grundlage geschaffen, um mögliche Übereinstimmungen und Unterschiede zwischen ihnen zu identifizieren. Dieser Schritt ist entscheidend, um einen reibungslosen und präzisen Übergang zwischen den Formaten zu ermöglichen und sicherzustellen, dass die Konkordanzen korrekt und zuverlässig erstellt werden können.

Das modifizierte Allegro-Neutralformat und das RIS-Format sind zwei bibliographische Datenformate, die speziell für die Darstellung bibliographischer Objekte in einem maschinenlesbaren Format entwickelt wurden.[15] Ein bibliographisches Objekt wird als eine Repräsentation eines Dokuments definiert, das in einem Verzeichnis gespeichert werden kann.[16] Diese Objekte enthalten spezifische Informationen, die innerhalb der vorgestellten Formate in dafür vorgesehenen Feldern mit eigenen Bezeichnungen gespeichert werden. Alle Felder, die zu einem bestimmten bibliographischen Objekt gehören, bilden gemeinsam einen Eintrag innerhalb der Formate.[17]

Die erlaubten Felder und ihre Nutzung, sowie die Abgrenzung von Einträgen voneinander, bilden die spezifischen Formate, deren Hauptziel es ist, die Verarbeitung und den Austausch bibliographischer Objekte zu ermöglichen.[18] Obwohl beide Formate dieses Ziel verfolgen, weisen sie doch unterschiedliche Ausprägungen und Strukturen auf, die jeweils für bestimmte Anwendungen und Systeme optimiert sind, wie aus den nachfolgenden Kapiteln hervorgeht.

2.1 Modifiziertes Allegro-Neutralformat

Das modifizierte Allegro-Neutralformat ist eine Variante des Allegro-Neutralformats. Letzteres wurde speziell für die Verwaltung von bibliothekarischen Datenbanken entwickelt.[19]

Es nutzt dreistellige alphanumerische Bezeichnungen für Felder, welche optionale Unterfelder enthalten können, die durch einzelne Buchstaben gekennzeichnet sind und

[15] Vgl. Neumann, E. (o.D.). *allegro-C: Das Neutralformat.* allegro. https://www.allegro-c-support.de/doku/neutral/. [Letzter Aufruf: 11.03.2024]. ; Vgl. The Thomson Corporation. (06.10.2011). *RIS format specifications.* Wayback Machine. https://web.archive.org/web/20120526103719/http://refman.com/support/risformat_intro.asp. [Letzter Aufruf: 11.03.2024].

[16] Vgl. Gödert, W., Lepsky, K., & Nagelschmidt, M. (2012). *Informationserschließung und automatisches Indexieren: Ein Lehr- und Arbeitsbuch.* S.120.

[17] Vgl. Neumann, E. (o.D.). *allegro-C: Das Neutralformat.* ; Vgl. The Tomson Corporation. (14.02.2004). *RIS format specifications: Tag Format.* Wayback Machine. https://web.archive.org/web/20110926022719/http://www.refman.com/support/risformat_fields_01.asp. [Letzter Aufruf: 11.03.2024].

[18] Vgl. ebd.

[19] Vgl. Neumann, E. (o.D.). *allegro-C: Das Neutralformat.*

zusätzliche Informationen zu den Hauptfeldern bieten.[20] Ein weiteres Merkmal des Formats ist seine flexible Struktur, die es ermöglicht, Felder in beliebiger Reihenfolge anzuordnen und mehrfach zu verwenden.[21] Darüber hinaus können Felder auch mehrere Inhalte enthalten, die durch Trennzeichen voneinander abgegrenzt sind, was die Vielseitigkeit bei der Datenverwaltung erhöht.[22]

Ein zusätzliches Merkmal dieses Formats ist seine vordefinierte Liste von Dokumenttypen. Diese Dokumenttypen sind durch ein- bis zweistellige Buchstabencodes gekennzeichnet und umfassen insgesamt 23 verschiedene Arten von Dokumenten.[23] Diese dienen dazu, die verschiedenen Arten von Materialien in der Datenbank zu klassifizieren und ermöglichen so eine gezielte Suche und Organisation der Informationen.

Das Allegro-Neutralformat wurde speziell für die Datenbank "Literatur zur Informationserschließung" angepasst, um deren Inhalt besser darstellen zu können.[24] Eine grundlegende Änderung betrifft den Aufbau seiner Einträge, die stets mit dem Feld "000" für die Eintrags-ID beginnen und durch drei Et-Zeichen abgeschlossen werden.[25] Es ist weiterhin die Darstellung mehrerer Inhalte in einem Feld mithilfe von Trennzeichen möglich.[26] Dies wird genauer im Kapitel "4 Untersuchung der Datenbank" untersucht.

Um eine klare Unterscheidung zwischen Feldbenennungen und ihren Inhalten zu gewährleisten, wird eine Trennung durch Doppelpunkte vorgenommen.[27] Im Gegensatz zum ursprünglichen Allegro-Neutralformat, das 277 verschiedene Felder umfasst, werden hier 30 verschiedene Felder genutzt.[28] Unterfelder werden nicht verwendet, was die Struktur vereinfacht.[29]

Einige Felder wurden umgewidmet, diese sind "50p", "50r", "53a", "520", "540", "590", "860", "875" und "890",[30] während andere Felder wie "50c", "50p", "60b", "60a", "60d", "60e", "60f", "60g", "60k", "60l", "60s", "60r" und "892" neu hinzugefügt wurden,[31] um verschiedene Verschlagwortungen und bibliothekarische Klassifikationen darstellen zu können. Insgesamt verfügt das modifizierte Allegro-Neutralformat über 43 Felder, die eine präzise und effiziente Datenorganisation innerhalb der Datenbank ermöglichen.[32]

---

[20] Vgl. ebd.

[21] Vgl. Neumann, E. (o.D.). *allegro-C: Das Neutralformat: Ausführliche Liste aller Felder.* allegro. https://www.allegro-c-support.de/doku/neutral/tab.htm. [Letzter Aufruf: 11.03.2024].

[22] Vgl. ebd.

[23] Vgl. ebd.

[24] Vgl. Lepsky, K. (05.2017). *Kategorienschema der Datenbank "Literatur zur Informationserschließung".*

[25] Vgl. "litie34.dbm" in den beigefügten Anlagen.

[26] Vgl. ebd.

[27] Vgl. ebd.

[28] Vgl. Lepsky, K. (05.2017). *Kategorienschema der Datenbank "Literatur zur Informationserschließung".*

[29] Vgl. ebd.

[30] Vgl. ebd.

[31] Vgl. ebd.

[32] Vgl. ebd.

Die Felder des modifizierten Allegro-Neutralformats und ihr jeweiliger Inhalt können folgender Tabelle entnommen werden:

Feldbenennung	**Inhalt**
000	Identnummer
030	Dokumenttyp
040	Erscheinungsjahr
050	Sprache
08I	ISBN
100	Sachtitel : Zusatztitel
200	Verfasser
21E	Herausgeber
270	Biografierte
411	Serientitel ; Zählung
510	Themenfeld
520	Wissenschaftsfach
530	Behandelte Form
540	Behandeltes Objekt
550	Behandeltes Land/Ort
590	Bibliothekarische Sparte
700	Quelle
810	Umfang
855	Ausgabevermerk
860	Impressum: Erscheinungsort : Verlag, Erscheinungsjahr
870	Abstract
875	Inhalt
880	Fußnote
50C	Compass
50L	LCSH (Library of Congress Subject Headings)
50P	Precis
50R	RSWK (Regeln für die Schlagwortkatalogisierung)
53A	Hilfsmittel
60B	BK (Basisklassifikation)
60A	ASB (Allgemeine Systematik für öffentliche Bibliotheken)
60D	DDC (Dewey Decimal Classification)

60E	Eppelsheimer
60F	SFB (Systematik für Bibliotheken)
60G	GHBS (Gesamthochschulbibliothekssystematik)
60K	KAB (Klassifikation für Allgemeinbibliotheken)
60L	LCC (Library of Congress Classification)
60S	SSD (Systematik der Stadtbibliothek Duisburg)
60R	RVK (Regensburger Verbundklassifikation)
890	Bildlink
892	Abbildung
900	Signatur
9ZC	Zugangsdatum
9ZE	Änderungsdatum

Tabelle 1: Felder des modifizierten Allegro-Neutralformats (Quelle: https://ixtrieve.fh-koeln.de/lehre/kategorienschema-litac.pdf)

Das modifizierte Allegro-Neutralformat weist eine andere Anzahl von Werten für Dokumenttypen auf als es das ursprüngliche Allegro-Neutralformat tut. Diese werden durch ein bis drei Buchstaben dargestellt und umfassen insgesamt 22 verschiedene Typen. Die Identifizierung dieser Typen kann durch Verweise im Feld "030" erfolgen, die mithilfe des regulären Ausdrucks "^030:.*(?:s\.|siehe).*$" abgerufen werden können. Darüber hinaus kann eine weitere Suche nach Sonderzeichen im Feld "030" den Dokumenttyp "?" hervorbringen, der durch den regulären Ausdruck "^030:.*\?.*$" bestätigt werden kann.

Die Werte des modifizierten Allegro-Neutralformats für Dokumenttypen können folgender Tabelle entnommen werden:

Wert	**Dokumenttyp**
a	Aufsatz
ag	Artikel
au	Konferenzschrift
b	Bibliographie
d	Dissertation
el	Elektronisches Dokument
fi	Mikroform
h	Anleitung
i	Nachschlagewerk/Informationsmittel
l	Loseblattsammlung
m	Buch

ms	Teil einer monographischen Reihe
n	Norm
p	Neudruck
pat	Patent
r	Bericht
s	Sammelwerk/Kongresspublikation/Buch
u	Skript/Unterrichtsmaterialien
vi	Video/Film
x	Hausarbeit/Diplomarbeit
z	Zeitschrift
?	Unbekannt

Tabelle 2: Dokumenttypen des modifizierten Allegro-Neutralformats (Quelle: "litie34.dbm" in den beigefügten Anlagen)

Das modifizierte Allegro-Neutralformat ist ein bibliothekarisches Datenformat, das eine Ausprägung bibliographischer Datenformate darstellt und entsprechend von Bibliotheken und Bibliothekssystemen genutzt wird.[33] In diesem Kontext stehen bibliothekarische Ansprüche im Vordergrund, die darauf abzielen, die Auffindbarkeit eines Dokuments innerhalb einer umfangreichen Sammlung zu gewährleisten. Dies erfordert, dass bibliographische Objekte eine detaillierte Beschreibung des repräsentierten Dokuments bieten, um dieses Ziel zu erreichen.[34] Außerdem sind bestimmte Felder vorhanden, die spezifisch bibliothekarische Informationen darstellen und somit eine effiziente Verwaltung und Recherche in Bibliothekssystemen ermöglichen.[35]

Dies manifestiert sich im modifizierten Allegro-Neutralformat in verschiedenen Sachverhalten. Zu diesem Zweck werden die sieben Felder "270", "510", "520", "530", "540", "550" und "590" für die Darstellung von Sachschlagworten bereitgestellt. Diese Felder dienen dazu, relevante Schlüsselwörter zu identifizieren und zu kategorisieren, um die Suche und den Zugriff auf die Inhalte zu erleichtern.

Des Weiteren stehen 14 Felder zur Verfügung, um bibliothekarische Klassifikationen und Notationen anzuzeigen. Hierbei handelt es sich um "50C", "50L", "50P", "50R", "60B", "60A", "60D", "60E", "60F", "60G", "60K", "60L", "60S" und "60R". Diese ermöglichen eine präzise Zuordnung und Organisation der bibliografischen Einträge nach verschiedenen Klassifizierungssystemen und Notationen. Ein weiteres Feld in diesem Kontext ist das Feld "900", das zur Darstellung von Signaturen verwendet wird, um die Lokalisierung von Medien in Bibliotheken darzustellen.

[33] Vgl. Neumann, E. (o.D.). *allegro-C: Das Neutralformat.*

[34] Vgl. ebd.

[35] Vgl. Lepsky, K. (05.2017). *Kategorienschema der Datenbank "Literatur zur Informationserschließung".*

2.2 RIS-Format

Das RIS-Format ist ein Referenzdatenformat, das in Literaturverwaltungssystemen genutzt wird.[36] Die Feldbenennungen in diesem Format bestehen aus zwei alphanumerischen Zeichen, die durch einen Bindestrich von ihren Inhalten getrennt sind.[37] Jeder Eintrag im RIS-Format muss mit dem Feld "TY" für den Dokumenttyp beginnen und mit einem leeren Feld "ER" enden[38]. Es gilt zusätzlich, dass alle Felder außer "TY" und "ER" optional sind und in beliebiger Reihenfolge angeordnet werden können.[39] Darüber hinaus können Felder mehrmals verwendet werden, um verschiedene Informationen zu erfassen.[40]

Es muss beachtet werden, dass zwei Versionen des RIS-Formats existieren, welche jeweils im Jahr 2001 und 2011 veröffentlicht wurden und teilweise unterschiedliche Felder und Feldnutzungen besitzen.[41] Im Rahmen dieser Arbeit wurde entschieden, das aktuellere Format aus dem Jahr 2011 zu nutzen. Insgesamt stellt diese Version 57 verschiedene Felder zur Verfügung.[42]

Die Felder des RIS-Formats und ihr jeweiliger Inhalt können folgender Tabelle entnommen werden:

Feldbenennung	**Inhalt**
TY	Dokumenttyp
ER	Kein Inhalt
AU	Geistiger Schöpfer
A2	Sekundärer Autor/Darsteller/Sponsor/Herausgeber/ Protokollführer/Produzent/Empfänger/Dateiname
A3	Drittautor/Herausgeber/Illustrator/Produzent/Verleger/ Betreuer
A4	Viertautor/Übersetzer/Sponsor/Finanzier/Darsteller/ Abteilung
AB	Abstract
AD	Autorenadresse
AN	Zugangsnummer

[36] Vgl. The Thomson Corporation. (06.10.2011). *RIS format specifications.*

[37] Vgl. The Tomson Corporation. (14.02.2004). *RIS format specifications: Tag Format.*

[38] Vgl. The Thomson Corporation. (07.07.2024). *RIS format specifications: Tag Definitions: Title and Reference Type Tags.* Wayback Machine. https://web.archive.org/web/20100726184137/http://www.refman.com/support/risformat_tags_01.asp. [Letzter Zugriff: 11.03.2024].

[39] Vgl. ebd.

[40] Vgl. ebd.

[41] Vgl. [larsgw], [dstillman], [adamsmith]. (12.04.2021-13.04.2021). *Ris specification*. Zotero Forums. https://forums.zotero.org/discussion/89035/ris-specification. [Letzter Aufruf: 11.03.2024].

[42] Vgl. [Aurimasv]. (12.04.2012). *Ris tag map: aurismasv/Translators Wiki.* GitHub. https://github.com/aurimasv/translators/wiki/RIS-Tag-Map. [Letzter Aufruf: 11.03.2024].

C1	Beliebig/Rechtlicher Hinweis/Besetzung/ PMID-Nummer/Autor/Zugehörigkeit/Abschnitt/ Publikationsort/Zeitraum/Begriff/Zeitraum/Zitationsjahr/ Abschnitt/Begriff/Regierungsbehörde/ Kontaktname/Größe/Musikformat/Sender-E-Mailadresse
C2	Beliebig/Danksagung/Publikationsjahr/ Beobachtungseinheit/Zitationsdatum/ Kongressnummer/Kontaktadresse/Gebiet/ Kompositionsform/Ausgabedatum/ Empfänger-E-Mailadresse/Berichtnummer
C3	Beliebig/Größe/Titelpräfix/Bandtitel/Datentyp/ PMC-Nummer/Kongresssitzung/Kontaktnummer/ Kontaktfaxnummer/Musikteil/ Designierte Staaten/Anwalt
C4	Beliebig/Gutachter/Datensatz/Zielgruppe/Referenz
C5	Beliebig/Format/Verpackungsmethode/Ausgabentitel/ Letztes Änderungsdatum/Finanzierungsnummer/ Begleitmaterial/Länge/Rechtsstatus/Verleger
C6	Beliebig/NIHMS-Nummer/PMC-Nummer/ CFDA-Nummer/Rechtsstatus/Ausgabe/Band
C7	Beliebig/Artikelnummer
C8	Beliebig
CA	Beschriftung
CN	Kennnummer
CT	Beschriftung
CY	Publikationsort/Stadt
DA	Erscheinungsdatum/Änderungsdatum/ Zugangsdatum/Frist/Inkrafttretungsdatum
DB	Datenbankname
DO	DOI (Digital Object Identifier)
DP	Datenbankanbieter
ET	Ausgabe/Epub-Datum/Publikationsdatum/Sitzung/ Klage des Obergerichts/Version/Beschreibung/ Internationale Patentklassifikation
IS	Ausgabe/Bandmenge
J2	Alternativtitel/Abgekürzter Titel
KW	Schlagwort
L1	Dateianhang
L4	Abbildung
LA	Sprache
LB	Label

M1	Nummer/Größe/Serienband/Computer/Ausgabe/ Kapitel/Status/Startseite/Zugangsdatum
M2	Startseite/Seitenmenge
M3	Typ/Rückwärtszitat/Form
N1	Notizen
NV	Bandmenge/Arbeitsumfang/Reporterkürzel/ Katalognummer/Nummer/Version/Häufigkeit/ US-Patentklassifikation/Serienband/Bandmenge
OP	Originalpublikation/Inhalt/Geschichte/ Versionsgeschichte/Ursprüngliche Förderungsnummer/ Prioritätennummer
PB	Verleger/Sponsor/Bibliothek/Archiv
PY	Jahr
RI	Untersuchter Gegenstand/Artikelnummer
RN	Forschungsnotizen
RP	Neudruckausgabe/Notizen
SE	Abschnitt/Nachrichtennummer/Seiten/Kapitel/ Einreichedatum/Veröffentlichungsdatum/Version/ Epub-Datum/Finanzierungsdauer/Abschnittsnummer/ Startseite/Internationale Patentnummer
SN	ISBN/ISSN/Berichtnummer/Dokumentnummer/ Patentnummer
SP	Seiten/Startseite/Beschreibung/Laufzeit
ST	Kurztitel
SV	Serienband
TI	Titel/Name
T2	Nebentitel/Zeitschrift/Quellentitel/Quellenname/Code/ Bildquellenprogramm/Komitee
T3	Dritttitel/Quellentitel/Quellenname/Institution/ Entscheidung/Legislatives Organ
TA	Übersetzter Autor
TT	Übersetzter Titel
UR	URL
VL	Band/Bildgröße/Angeforderter Betrag/Lagerung/ Patentversionsnummer/Regelnummer/Codenummer/ Akademischer Grad/Zugangsdatum
Y2	Zugangsdatum/Inkrafttretungsdatum

Tabelle 3: Felder des RIS-Formats (Quelle: https://github.com/aurimasv/translators/wiki/RIS-Tag-Map)

Innerhalb des RIS-Formats werden Werte verwendet, die Dokumenttypen repräsentieren.[43] Diese bestehen aus Großbuchstaben, die meistens Kürzel des bezeichneten Dokumenttyps darstellen und umfassen insgesamt 56 verschiedene Typen.[44] Die Werte des RIS-Formats für Dokumenttypen können folgender Tabelle entnommen werden:

Wert	**Dokumenttyp**
ABST	Abstract
INPR	Inprint
JFULL	Zeitschrift
JOUR	Artikel
AGGR	Aggregierte Datenbasis
ANCIENT	Antiker Text
ART	Kunstwerk
ADVS	Audiovisuelles Material
SLIDE	Präsentationsfolie
SOUND	Geräusch
VIDEO	Video
BILL	Rechnung
BLOG	Blog
BOOK	Buch
CHAP	Buchteil
CASE	Fall
CTLG	Katalog
CHART	Graphik
EQUA	Gleichung
CLSWK	Klassisches Werk
COMP	Computerprogramm
DATA	Datei
CPAPER	Tagungspapier
CONF	Tagungsband
DICT	Wörterbuch
EDBOOK	Herausgegebenes Buch

[43] Vgl. The Thomson Corporation. (07.07.2024). *RIS format specifications: Tag Definitions: Title and Reference Type Tags.*

[44] Vgl. [Aurimasv]. (12.04.2012). *Ris tag map: aurismasv/Translators Wiki.*

EJOUR	Elektronischer Artikel
EBOOK	E-Book
ECHAP	Elektronischer Buchteil
ENCYC	Enzyklopädie
FIGURE	Abbildung
MPCT	Fernsehausstrahlung
GOVDOC	Regierungsdokument
GRANT	Genehmigung
HEAR	Gehörtes
LEGAL	Gesetz
MGZN	Magazinartikel
MANSCPT	Manuskript
MAP	Karte
MUSIC	Musik
NEWS	Zeitungsartikel
DBASE	Datenbank
MULTI	Multimediales Material
PAMP	Pamphlet
PAT	Patent
PCOMM/ICOMM	Persönliche Kommunikation
RPRT	Bericht
SER	Reihe
STAND	Standard
STAT	Statute
THES	Thesis
UNPB	Unveröffentlichte Arbeit
ELEC	Webseite

Tabelle 4: Dokumenttypen des RIS-Formats (Quelle: https://github.com/aurimasv/translators/wiki/RIS-Tag-Map)

Da das RIS-Format ein Referenzdatenformat ist, das für bibliographische Referenzen in Bibliographien und Zitationen verwendet wird,[45] stellt es eine Ausprägung von bibliographischen Datenformaten dar. Dabei ist der Aufbau solcher bibliographischen Referenzen stark vom Dokumenttyp des repräsentierten Dokuments abhängig. Aus

[45] Vgl. The Thomson Corporation. (06.10.2011). *RIS format specifications.*

diesem Grund ist die formale Beschreibung von Dokumenten von besonderer Bedeutung.[46] Im RIS-Format spiegelt sich diese Anforderung in verschiedenen Aspekten wider.

Von den insgesamt 57 Feldern im RIS-Format hängt die Verwendung von 31 Feldern vom Dokumenttyp des Eintrags ab. Dies unterstreicht die Notwendigkeit, den Dokumenttyp genau zu bestimmen, um die entsprechenden Felder korrekt zu verwenden. Außerdem bietet das RIS-Format mit 56 Dokumenttypen nahezu so viele Dokumenttypen wie es Felder hat. Trotz der Vielzahl von Feldern und Dokumenttypen gibt es lediglich drei Felder im RIS-Format, die zur inhaltlichen Beschreibung des Dokuments dienen, nämlich "AB", "KW" und "RI".

[46] Vgl. Bertram, J. (2019). *Abschlussarbeiten in der Bibliotheks-und Informationswissenschaft.* De Gruyter Saur. S.141f.

3 Konkordanz zwischen den Formaten

Um die Konvertierung der Datenbank durchführen zu können, ist es notwendig, Konkordanzen zwischen dem modifizierten Allegro-Neutralformat und dem RIS-Format zu erstellen. Diese dienen als Grundlage, um festzustellen, welche Felder auf welche Weise übersetzt werden müssen, wodurch sichergestellt wird, dass die Datenbank bei der Konvertierung korrekt und einheitlich strukturiert ist.

Dabei existieren einige Differenzen zwischen dem modifizierten Allegro-Neutralformat und dem RIS-Format, die aus ihren jeweiligen Verwendungszwecken resultieren, wie im vorherigen Kapitel erläutert wurde. Diese Unterschiede haben direkte Auswirkungen auf die Erstellung einer Konkordanz zwischen den beiden Formaten.

3.1 Dokumenttypen

Hier wird eine Konkordanz zwischen den verwendeten Dokumenttypen des modifizierten Allegro-Neutralformats und des RIS-Formats erstellt.

In dem modifizierten Allegro-Neutralformat sind insgesamt elf Dokumenttypen vorhanden, die direkt durch entsprechende Dokumenttypen im RIS-Format ausgedrückt werden können. Dies ermöglicht eine unmittelbare Übersetzung zwischen den beiden Formaten. Zu diesen elf Dokumenttypen gehören "ag", "au", "d", "m", "ms", "pat", "r", "s", "vi", "x" und "z".

Besonders hervorzuheben ist der Dokumenttyp "a", da er sowohl Zeitungsartikel als auch Aufsätze in Sammelwerken repräsentiert. Daher kann diesem Dokumenttyp sowohl der RIS-Dokumenttyp "JOUR" für Zeitschriftenartikel, als auch "CHAP" für Aufsätze in Sammelwerken zugeordnet werden.

Es existieren außerdem fünf Dokumenttypen im modifizierten Allegro-Neutralformat, die zwar in einen RIS-Dokumenttyp übersetzt werden können, jedoch ohne dass dies eine präzise Entsprechung darstellt. Dies liegt daran, dass einige der Dokumenttypen nicht genau mit den RIS-Dokumenttypen übereinstimmen. Stattdessen werden hinreichende Annäherungen gewählt, um die Informationen angemessen zu repräsentieren. Zu den betroffenen Dokumenttypen gehören "fi", welches als "ADVS" übersetzt wird, "n" als "STAT", "p" als "JOUR", "u" als "SLIDE" und schließlich "?" als "GEN". Diese Auswahl reflektiert die Notwendigkeit, die spezifischen Eigenschaften und Inhalte der Originaldokumente bestmöglich in das RIS-Format zu integrieren, obwohl eine exakte Übersetzung aufgrund der Unterschiede zwischen den Formaten nicht immer möglich ist.

Zusätzlich gibt es vier Dokumenttypen, die in den RIS-Dokumenttyp "BOOK" übersetzt werden, obwohl diese Übersetzung nicht akkurat ist. Die betroffenen Dokumenttypen sind "b", "h", "i" und "l". Diese Situation ergibt sich aufgrund der Schwierigkeiten bei ihrer Darstellung mit den RIS-Dokumenttypen. Der Dokumenttyp "BOOK" wurde schließlich

gewählt, da dieser aufgrund seiner einfachen Verwendung von Feldern als Standarddokumenttyp genutzt werden kann.

Ein ähnlicher Fall liegt beim Dokumenttyp des modifizierten Allegro-Neutralformats "el" vor, der in den RIS-Dokumenttyp "EBOOK" übersetzt wird. Diese Entscheidung beruht auf der Tatsache, dass diese Dokumente als elektronische Ressourcen klassifiziert sind, was die Wahl des RIS-Dokumenttyps "EBOOK" als angemessen erscheinen lässt.

Die komplette Konkordanz zwischen den Dokumenttypen des modifizierten Allegro-Neutralformats und des RIS-Formats kann folgender Tabellen entnommen werden:

Allegro-Neutralformat	**RIS-Format**
a (Artikel)	JOUR
a (Aufsatz)	CHAP
ag (Artikel)	JOUR
au (Konferenzschrift)	CONF
b (Bibliographie)	BOOK
d (Dissertation)	THES
el (Elektronisches Dokument)	EBOOK
fi (Mikroform)	ADVS
h (Anleitung)	BOOK
i (Nachschlagewerk/Infomittel)	BOOK
l (Loseblattsammlung)	BOOK
m (Monographie)	BOOK
ms (Teil einer monographischen Reihe)	BOOK
n (Norm)	STAT
p (Preprint)	JOUR
pat (Patent)	PAT
r (Bericht)	RPRT
s (Sammelwerk)	BOOK
u (Skript/Unterrichtsmaterialien)	SLIDE
vi (Video/Film)	VIDEO
x (Hausarbeit/Diplomarbeit)	THES
z (Zeitschrift)	JFULL
? (Unbekannt)	GEN

Tabelle 5: Konkordanz zwischen Dokumenttypen

3.2 Felder

Im nächsten Schritt wird eine Konkordanz zwischen den Feldern des modifizierten Allegro-Neutralformats und des RIS-Formats erstellt. Es ist jedoch zu beachten, dass die Bedeutung einiger Felder des RIS-Formats von dem Dokumenttyp ihres Eintrags abhängt. Daher muss bei der späteren Übersetzung vom modifizierten Allegro-Neutralformat in das RIS-Format auf diese Abhängigkeit geachtet werden. Eine detaillierte Untersuchung des Umgangs mit spezifischen Feldern in diesem Kontext findet im Kapitel "4 Untersuchung der Datenbank" statt.

Im modifizierten Allegro-Neutralformat gibt es 13 Felder, die direkt durch Felder im RIS-Format ausgedrückt werden können. Dabei handelt es sich um die Felder "030", "040", "050", "08I", "200", "21E", "411", "810", "855", "870", "890", "892" und "9ZC". Dies setzt aber voraus, dass der entsprechende Dokumenttyp diese Interpretation des Feldes unterstützt.

Außerdem sind drei Felder identifiziert worden, welche die Möglichkeit bieten, mehrere verschiedenartige Informationen zu enthalten. Diese Informationen können jeweils durch Felder im RIS-Format ausgedrückt werden, sofern der entsprechende Dokumenttyp diese Interpretation des Feldes vorsieht. Eine adäquate Übersetzung dieser Felder in das RIS-Format erfordert daher eine Trennung der verschiedenen Inhalte, um eine korrekte Darstellung zu gewährleisten. Dies betrifft die Felder "100", "700" und "860".

Es existieren zwei Felder, die in ein RIS-Feld umgewandelt werden, obwohl diese Übersetzung nicht akkurat ist. Dies liegt an der Schwierigkeit, diese Felder präzise mit RIS-Feldern darzustellen, weshalb die gewählten RIS-Felder als hinreichende Annäherungen betrachtet werden. Die betroffenen Felder sind das Feld "000", welches in das Feld "CN" übersetzt wird und das Feld "900", das in das Feld "AN" umgewandelt wird.

Dazu umfasst das modifizierte Allegro-Neutralformat sieben Felder, die jeweils mit verschiedenen Sachschlagwörtern verbunden sind, nämlich "270", "520", "520", "530", "540", "550" und "590". Obwohl diese Felder nicht präzise in das RIS-Format übertragen werden können, ist dennoch eine jeweilige Übersetzung in das RIS-Feld "KW" möglich, welches ein allgemeines Feld für Schlagwörter darstellt. Diese Übersetzung führt zu einer Unschärfe und kann somit zu einem Informationsverlust führen. Wodurch der Informationsverlust akzeptabel wird, ist der Erhalt der grundsätzlichen Funktion der Schlagwörter.

Von besonderem Interesse sind 17 Felder, die nicht im RIS-Format dargestellt werden können. Um sicherzustellen, dass die Inhalte dieser Felder nicht verloren gehen, werden sie in das RIS-Feld "N1" übersetzt, welches Notizen beschreibt. Zusätzlich wird jeder Inhalt mit einer spezifischen Bezeichnung eingeführt, die durch einen Doppelpunkt vom eigentlichen Inhalt getrennt ist. Zum Beispiel wird ein solcher Eintrag wie folgt aussehen: "N1 - DDC: 155.95". Auf diese Weise bleiben die Informationen dieser Felder weiterhin für Untersuchungen und Verarbeitungen verfügbar.

Diese spezielle Vorgehensweise betrifft insbesondere die Felder, die bibliothekarische Klassifikationen und Notationen darstellen, nämlich "50C", "50L", "50P", "50R", "60B", "60A", "60D", "60E", "60F", "60G", "60K", "60L", "60S" und "60R". Darüber hinaus sind auch die Felder "875", "880" und "53A" von dieser Übersetzung betroffen. Durch diese Maßnahme bleibt die Integrität und Verfügbarkeit der Informationen in den Feldern gewährleistet, auch wenn sie nicht direkt dargestellt werden können.

Eine Besonderheit stellt das Feld "9ZE" dar. Dieses Feld lässt sich ausschließlich bei den Dokumententypen "EBOOK", "BLOG" und "ELEC" direkt in das RIS-Format übertragen.[47] Jedoch ist der Informationsgehalt dieses Feldes minimal, da es lediglich Aussagen über den Eintrag selber und nicht das repräsentierte Objekt enthält. Aus diesem Grund wurde entschieden, dieses Feld nicht zu übersetzen und den damit verbundenen Informationsverlust hinzunehmen.

Die komplette Konkordanz zwischen den Feldern des modifizierten Allegro-Neutralformats und des RIS-Formats kann folgender Tabellen entnommen werden:

Allegro-Neutralformat	**RIS-Format**
000: Identnummer	CN - Kennnummer
030: Dokumenttyp	TY - Dokumenttyp
040: Erscheinungsjahr	PY – Jahr
050: Sprache	LA - Sprache
08I: ISBN	SN - ISBN/ISSN/Berichtnummer/Dokumentnummer/ Patentnummer
100: Sachtitel : Zusatztitel	TI - Titel/Name
	T2 - Nebentitel/Zeitschrift/Quellentitel/ Quellen-name/Code/Bildquellenprogramm/Komitee
200: Verfasser	AU – Geistiger Schöpfer
21E: Herausgeber	A2 - Sekundärer Autor/Darsteller/Sponsor/ Herausgeber/Protokollführer/Produzent/Empfänger/ Dateiname
270: Biografierte	KW - Schlagwort
411: Serientitel ; Zählung	T3 - Dritttitel/Quellentitel/Quellenname/Institution/ Entscheidung/Legislatives Organ
510: Themenfeld	KW - Schlagwort
520: Wissenschaftsfach	KW - Schlagwort
530: Behandelte Form	KW - Schlagwort
540: Behandeltes Objekt	KW - Schlagwort
550: Behandeltes Land/Ort	KW - Schlagwort

[47] Vgl. [Aurimasv]. (12.04.2012). *Ris tag map: aurismasv/Translators Wiki.*

590: Bibliothekarische Sparte	KW - Schlagwort
700: Quelle	T2 - Nebentitel/Zeitschrift/Quellentitel/ Quellenname/Code/Bildquellenprogramm/Komitee
	VL - Band/Bildgröße/Angeforderter Betrag/Lagerung/ Patentversionsnummer/Regelnummer/Codenummer/ Akademischer Grad/Zugangsdatum
	IS - Ausgabe/Bandmenge
	DA - Erscheinungsdatum/Änderungsdatum/ Zugangsdatum/Frist/Inkrafttretungsdatum
	UR - URL
	SP - Seiten/Startseite/Beschreibung/Laufzeit
810: Umfang	SP - Seiten/Startseite/Beschreibung/Laufzeit
855: Ausgabevermerk	ET - Ausgabe/Epub-Datum/Publikationsdatum/ Sitzung/Klage des Obergerichts/Version/Beschreibung/ Internationale Patentklassifikation
860: Impressum: Erscheinungsort : Verlag, Erscheinungsjahr	AD - Autorenadresse
	CY - Publikationsort/Stadt
	PB - Verleger/Sponsor/Bibliothek/Archiv
	PY – Jahr
870: Abstract	AB - Abstract
875: Inhalt	N1 - Notizen
880: Fußnote	N1 - Notizen
50C: Compass	N1 - Notizen
50L: LCSH	N1 - Notizen
50P: Precis	N1 - Notizen
50R: RSWK	N1 - Notizen
53A: Hilfsmittel	N1 - Notizen
60B: BK	N1 - Notizen
60A: ASB	N1 - Notizen
60D: DDC	N1 - Notizen
60E: Eppelsheimer	N1 - Notizen
60F: SFB	N1 - Notizen
60G: GHBS	N1 - Notizen
60K: KAB	N1 - Notizen
60L: LCC	N1 - Notizen
60S: SSD	N1 - Notizen

60R: RVK	N1 - Notizen
890: Bildlink	L1 – Dateianhang
892: Abbildung	L4 - Abbildung
900: Signatur	AN - Zugangsnummer
9ZC: Zugangsdatum	Y2 - Zugangsdatum/Inkrafttretungsdatum
9ZE: Änderungsdatum	-

Tabelle 6: Konkordanz zwischen Feldern

4 Untersuchung der Datenbank

Zunächst folgt eine Analyse der Inhalte der Datenbank "Literatur zur Informationserschließung". Eine gründliche Untersuchung ist nötig, um festzustellen, ob Felder mehrere Inhalte aufnehmen können, welche spezifischen Inhalte auftreten können und wie Felder mit mehreren Inhalten voneinander abgrenzt werden. Diese ist von entscheidender Bedeutung für die spätere Konvertierung der Datenbank, da sie Einblicke darüber liefert, wie die Inhalte am besten aufgeteilt werden können. Insbesondere geht es darum, festzustellen, wie Felder mit mehreren Inhalten effizient in das RIS-Format konvertiert werden können. Dabei werden nur Inhalte der Datenbank untersucht, die für die Konvertierung relevant sind.

Die Untersuchung basiert auf der Anwendung regulärer Ausdrücke, die es ermöglichen, bestimmte Muster innerhalb von Datensätzen zu identifizieren. Dabei werden zunächst alle Felder isoliert, wobei für jedes Feld passende reguläre Ausdrücke verwendet werden, die die zu untersuchenden Zeichen ergänzen. Ein zentraler Aspekt dieser Methode besteht darin, nach potenziellen Trennzeichen innerhalb der Felder zu suchen, um festzustellen, wie diese in der Praxis verwendet werden. Dies wird durch das systematische Absuchen nach Leerzeichen, senkrechten Strichen, Kommata, Punkten, Doppelpunkten, Semikola, Bindestrichen, Pluszeichen und Schrägstrichen erreicht.

Ein konkretes Beispiel für die Anwendung dieser Methode ist der reguläre Ausdruck "^050:.*\|.*$", der darauf abzielt, Felder mit der Bezeichnung "050" zu identifizieren, die einen senkrechten Strich enthalten. Hierbei werden sowohl der Feldname als auch das zu suchende Zeichen entsprechend angepasst. Die Ergebnisse dieser Suche geben Aufschluss darüber, ob und wie diese Zeichen tatsächlich als Trennzeichen genutzt werden.

Es ist wichtig anzumerken, dass Abweichungen von dieser standardisierten Vorgehensweise dokumentiert werden. Sollten andere reguläre Ausdrücke verwendet werden, die von der beschriebenen Methodik abweichen, so werden diese entsprechend vermerkt. Auf diese Weise wird eine transparente und nachvollziehbare Analyse ermöglicht, die die Grundlage für weiterführende Untersuchungen bildet.

Die Analyse von Verbindungen zwischen verschiedenen Bereichen oder Inhalten in Einträgen erfolgt mithilfe des Pythonskripts. Reguläre Ausdrücke, die für die Untersuchung dieser Feldverknüpfungen verwendet werden, sowie ihr Kontext werden entsprechend angegeben. Im Kapitel "6 Ergebnis der Konvertierung" wird die Korrektheit der Annahmen, die aus den Untersuchungen resultieren, im Rahmen der Kontrolle des Konvertierungsprozesses geprüft.

In den Feldern "000", "040", "870", "892" und "9ZC" können keine Trennzeichen festgestellt werden. Es wird daher angenommen, dass dieses Feld keine mehrfachen Inhalte enthalten. Folglich bleibt der Inhalt bei der Umwandlung in das RIS-Format unverändert.

Da der Inhalt des Feldes "030" der Dokumenttyp des entsprechenden Eintrags ist und somit anhand der erarbeiteten Konkordanz für Dokumenttypen konvertiert werden soll, wird der Inhalt dieses Feldes auch auf Übereinstimmung mit den Dokumenttypen des modifizierten Allegro-Neutralformats untersucht. Der Text enthält entweder einen Wert oder mehrere Werte des modifizierten Allegro-Neutralformats für einen bestimmten Dokumenttyp. Wenn es sich um mehrere Werte handelt, sind sie durch senkrechte Striche voneinander getrennt. Alternativ kann der Text auch eine Verweisung auf einen Wert des modifizierten Allegro-Neutralformats für einen Dokumenttyp enthalten. Diese Verweisungen beinhalten immer die Zeichenketten "siehe" oder "s.".[48]

Da bei Einträgen mit Verweisen das Feld "100" fehlt, wird davon ausgegangen, dass diese keine eigenständigen bibliographischen Objekte darstellen.[49] In solchen Fällen wird der gesamte Eintrag nicht in das RIS-Format übersetzt. Wenn ein Eintrag mehrere Dokumenttypen aufweist, wird nur der zuerst genannte Dokumenttyp übernommen, um mögliche Konflikte zwischen verschiedenen Typen zu vermeiden. Dabei wird bewusst akzeptiert, dass Informationen verloren gehen können.

In den Feldern "050", "08I", "200", "21E", "270", "510", "520", "530", "540", "550", "590", "890" und "900" können vertikale Striche als Trennzeichen identifiziert werden. Diese Striche dienen dazu, mehrere Inhalte der gleichen Art voneinander zu trennen. Während der Konvertierung in das RIS-Format wird der Inhalt des Feldes anhand dieser Trennzeichen aufgeteilt, bleibt aber ansonsten unverändert.

In dem Feld "100" können Doppelpunkte als Trennzeichen identifiziert werden. Diese dienen dazu, den Haupttitel von dem Zusatztitel zu trennen. Eine Ausnahme stellen URLs dar, welche auch in Titeln vorkommen können und immer Doppelpunkte enthalten, daher stellen Doppelpunkte in diesen Fällen keine Trennzeichen dar.[50]

Nur in wenigen Fällen erfolgt eine Trennung des Zusatztitels vom Haupttitel anhand eines Trennzeichens, da das RIS-Feld "T2" nur bei wenigen Dokumenttypen einen Zweittitel beschreibt. Wenn der Dokumenttyp des Eintrags "?" ist, wird der Zweittitel durch das Trennzeichen bestimmt, da das RIS-Feld "T2" bei der Übersetzung in den RIS-Dokumenttyp "GEN" den Zusatztitel angibt. Diese Regel gilt jedoch nicht, wenn der Titel eine URL ist. In allen anderen Situationen bleibt der Inhalt bei der Konvertierung in das RIS-Format unverändert.

In dem Feld "411" können Semikola und Kommata als Trennzeichen identifiziert werden. Diese dienen dazu, den Serientitel von dessen Zählung zu trennen. Während der Konvertierung in das RIS-Format bleibt der Inhalt des Feldes jedoch unverändert, da Kommata und Semikola in diesem Feld nicht nur als Trennzeichen, sondern teilweise

[48] Der reguläre Ausdruck "^030:.+$" findet 43.438 Ergebnisse, was bedeutet, dass das Feld "030" in 43.438 Einträgen vorhanden ist. Die regulären Ausdrücke. "^030:(?:\w{1}|\w{2}|\w{3}|\?)$", "^030:(?:\w{1}|\w{2}|\w{3}|\?)\|(?:\w{1}|\w{2}|\w{3}|\?)$", "^030:(?:\w{1}|\w{2}|\w{3})\|(?:\w{1}|\w{2}|\w{3})\|(?:\w{1}|\w{2}|\w{3})$" und "^030:.*(?:s\.|siehe).*$" decken alle diese ab.

[49] Das Pythonskript findet unter Angabe des regulären Ausdrucks "(?:s\.|siehe)" für das Feld "030" und eines leeren regulären Ausdrucks für das Feld "100" keine gemeinsamen Treffer

[50] Kann mit dem regulären Ausdruck "^100:.*https?://.+$" ermittelt werden

auch in den Serientiteln verwendet werden und eine zuverlässige Trennung der Inhalte somit nicht möglich ist. Der Erhalt dieser Informationen wird hier also zulasten der Qualität der entsprechenden RIS-Einträge beschlossen.

In Feld "700" werden verschiedene Inhalte in unterschiedlichen Konstellationen erfasst. Dies umfasst sowohl den Titel einer Quelle als auch eine URL, wobei diese durch einen Punkt abgeschlossen wird, wenn weitere Bestandteile im Feld vorhanden sind. Eine Bandangabe kann ebenfalls enthalten sein und wird üblicherweise mit "Vol.", "Nr." oder "Bd." eingeleitet, wobei dies nicht immer der Fall ist. Auch hier folgt ein Komma, wenn weitere Bestandteile im Feld vorhanden sind.

Des Weiteren kann eine Datumsangabe im Format "TT.MM.JJJJ" erscheinen, wobei Platzhalter für einzelne Bestandteile und Bindestriche für Datumsspannen verwendet werden können. Auch hier schließt ein Komma den Eintrag ab, wenn weitere Bestandteile im Feld vorhanden sind. Eine Ausgabenangabe, die mit "H.", "no." oder "nos." eingeleitet wird, kann ebenfalls erscheinen und wird gleichermaßen mit einem Komma abgeschlossen, wenn weitere Bestandteile im Feld vorhanden sind.

Die Seitenangabe wird meist mit "s." oder "S." eingeleitet und enthält in der Regel einen Bindestrich, Kommata oder Pluszeichen, um mehrere Seiten anzugeben. Zudem können Herausgeber-, Bearbeiter- oder Übersetzerangaben vorkommen, die oft mit "Hrsg.:", "Bearb.:", "ed." oder "ed. by" beginnen. Abschließend kann auch ein Ausgabevermerk oder eine URL, die in eckigen Klammern eingeschlossen ist, enthalten sein.

Bei der Konvertierung in das RIS-Format werden verschiedene Elemente des Inhalts auf unterschiedliche Weise verarbeitet. Der Titel der Quelle wird anhand des abschließenden Punktes erkannt. Falls kein Punkt vorhanden ist, wird davon ausgegangen, dass nur der Titel der Quelle im Feld vorhanden ist. URLs werden durch das Vorhandensein der Zeichenketten "http://" oder "https://" erkannt.

Band- und Ausgabeangaben werden entweder anhand ihrer einleitenden Zeichenketten identifiziert oder, wenn keine solche Zeichenkette vorhanden ist, wird angenommen, dass es sich um eine Bandangabe handelt. Diese Angaben werden außerdem durch Kommata von anderen Inhalten abgegrenzt. Datumsangaben werden aufgrund ihres klaren Aufbaus erkannt, während Seitenangaben anhand ihrer einleitenden Zeichenkette und Struktur erkannt werden können.

Herausgeber-, Bearbeiter-, Übersetzer- und Ausgabeangaben hingegen können nicht zuverlässig voneinander und von anderen Inhalten getrennt werden, da ihre Strukturen zu heterogen sind und nicht immer klar voneinander abgegrenzt werden können. So unterscheiden sich Trennzeichen zwischen diesen, Trennzeichen werden teilweise in den Angaben selber verwendet und Bezeichnungen für die Funktionen von Personen unterscheiden sich oder fehlen ganz. Um dennoch zu vermeiden, dass diese Informationen verloren gehen, werden sie im Feld "T2" des RIS-Formats ohne Trennung zusammengeführt. Dies stellt einen Kompromiss dar um Informationsverlust zu vermeiden, auch wenn dies zu einer inkorrekten Nutzung des Feldes führt.

Das Feld "700" wird nur in Verbindung mit dem Feld "860" bei Aufsätzen in Sammelwerken verwendet. Dies ermöglicht die Übersetzung des Dokumenttyps "a" im modifizierten Allegro-Neutralformat in die richtigen RIS-Dokumenttypen "JOUR" oder "CHAP". Diese Annahme basiert auf der intellektuellen Untersuchung von 25 zufällig ausgewählten Einträgen mit der Feldbenennung "700".[51]

Im Feld "810" können Kommata und Pluszeichen als Trennzeichen identifiziert werden. Diese dienen dazu, nicht zusammenhängende Seitenangaben voneinander zu trennen oder Seitenangaben mit unterschiedlichen Zählungen wie römischer und arabischer Zählung. Zusätzlich trennen Pluszeichen Zusatzmaterialien von den Seitenangaben. Der Inhalt des Feldes "810" bleibt unverändert, wenn er in das RIS-Format konvertiert wird.

In 440 Fällen tritt das Feld "810" in Einträgen auf, in denen das Feld "700" ebenfalls Seitenangaben enthält.[52] Dies kann zu Konflikten bei der Übersetzung in das RIS-Feld "SP" führen, da beide Felder Seitenangaben enthalten. Um solche Konflikte zu vermeiden, wird in diesen Fällen nur die Seitenangabe aus dem Feld "700" in das RIS-Feld "SP" übersetzt.

Im Feld "855" können Punkte, Semikola und Pluszeichen als Trennzeichen identifiziert werden. Die Punkte dienen dazu, mehrere Ausgabevermerke voneinander zu trennen, werden jedoch auch als Abkürzungs- und Satzzeichen verwendet. Semikola werden ebenfalls verwendet, um mehrere Ausgabevermerke zu trennen, aber auch um denselben Ausgabevermerk in verschiedenen Sprachen sowie Zusatzinformationen vom Ausgabevermerk zu unterscheiden. Pluszeichen dienen ebenfalls der Trennung von diesen und finden auch Verwendung in Aufzählungen.

Bei der Konvertierung in das RIS-Format wird der Inhalt des Ausgabevermerks anhand von Semikola aufgeteilt, während er ansonsten unverändert bleibt. Die anderen Trennzeichen haben gleichzeitig andere Funktionen und würden den Inhalt unzuverlässig trennen, weshalb diese bei der Konvertierung nicht berücksichtigt werden.

Im Feld "860" können verschiedene Informationen enthalten sein. Das kann einerseits der Erscheinungsort ohne weitere Angaben sein. Andererseits kann auch der Erscheinungsort zusammen mit dem Verlag aufgeführt werden, wobei beide durch einen Doppelpunkt voneinander getrennt sind. Ebenso kann das Feld lediglich das Erscheinungsjahr enthalten. Des Weiteren ist es möglich, dass mehrere Verlagsorte vorhanden sind, die durch Semikola voneinander getrennt werden.

Bei der Konvertierung in das RIS-Format wird der Inhalt anhand der Trennzeichen aufgeteilt, bleibt aber ansonsten unverändert. Ein wesentliches Unterscheidungsmerkmal ist das Vorhandensein eines Doppelpunkts. Ist ein Solcher vorhanden, wird davon ausgegangen, dass sowohl ein Erscheinungsort als auch ein

[51] Es wurde der reguläre Ausdruck "^700:.+$" verwendet, aus deren Ergebnismenge zufällige Einträge ausgewählt wurden. Die gewählten Einträge können den beigefügten Anlagen in dem Ordner "test" als "aufsätze.txt" entnommen werden.

[52] Das Pythonskript findet unter Angabe des regulären Ausdrucks "(?:(?:\s*S\.\s*(?:\d|x|\?|*|\s*)+(?:\s*(?:\-|\+)\s*(?:\d|x|\?|*|\s*)+)?)|(?:\s*(?:\d|x|\?|*|\s*)+(?:\s*(?:\-|\+)\s*(?:\d|x|\?|*|\s*)+)\s*S(?!\w)))" für das Feld 700 und eines leeren regulären Ausdrucks für das Feld "810" 440 gemeinsamen Treffer

Verlag angegeben sind. Fehlt jedoch der Doppelpunkt, wird angenommen, dass entweder nur der Erscheinungsort oder nur das Erscheinungsjahr angegeben ist. In diesem Fall kann das Erscheinungsjahr jedoch durch seine Struktur, die aus Zahlen besteht, von Erscheinungsorten unterschieden werden.

Das Feld "860" wird nur in Verbindung mit dem Feld "700" bei Aufsätzen in Sammelwerken verwendet. Dies ermöglicht die Übersetzung des Dokumenttyps "a" im modifizierten Allegro-Neutralformat in die richtigen RIS-Dokumenttypen "JOUR" oder "CHAP". Diese Annahme basiert auf der selben intellektuellen Untersuchung von 25 zufällig ausgewählten Einträgen, die bereits für das Feld "700" stattgefunden hat.

In den Feldern "875", "880", "50C", "50L", "50P", "50R", "53", "60B", "60A", "60D", "60E", "60F", "60G", "60K", "60L", "60S", "60R", "875" und "880" wird, wie bereits im Kapitel "3.2 Felder" besprochen, der Inhalt durch Hinzufügen einer einleitenden Zeichenkette zur Beschreibung des Feldinhalts ergänzt und dann in das RIS-Feld "N1" übersetzt. Da dieses Vorgehen lediglich der Archivierung des Feldes dient, wird der Inhalt nicht verändert.

Eine Ausnahme tritt bei den Feldern "875" und "880" auf, wenn diese eine URL enthalten, was durch das Vorhandensein der Zeichenketten "http://" und "https://" erkennbar ist. In diesem Fall wird die URL in das RIS-Feld "UR" übersetzt, und der restliche Inhalt des Feldes wird verworfen, da er wahrscheinlich keine relevanten Informationen außer der URL mehr enthält.

5 Programm zur Konvertierung der Datenbank

In diesem Kapitel wird das Programm zur Konvertierung der Datenbank "Literatur zur Informationserschließung" detailliert beschrieben. Ziel dieses Programms ist es die angestrebte Konvertierung effektiv umzusetzen und um dieses zu erreichen dienen die Konkordanzen zwischen dem modifizierten Allegro-Neutralformat und dem RIS-Format als Basis. Darüber hinaus dient die Untersuchung der Datenbank als eine weitere Grundlage die in die Entwicklung des Konvertierungsprogramms einfließt. Die Zusammenführung dieser Elemente ermöglicht eine präzise und zuverlässige Realisierung der bisher besprochenen Vorgehensweisen und damit auch die Umwandlung der Datenbank in das RIS-Format.

5.1 Beschreibung

Die Beschreibung der Funktionsweise des Programms folgt unabhängig von spezifischen Programmiersprachen. Dieser Ansatz zielt darauf ab, sich auf das allgemeine Verständnis des Konvertierungsprozesses zu konzentrieren, indem die grundlegenden Mechanismen und Abläufe des Programms erläutert werden.

Der Konvertierungsprozess in dem Programm wird zunächst durch ein Aktivitätsdiagramm veranschaulicht, das den Ablauf gemäß der Unified Modeling Language, auch als UML abgekürzt, definiert. Diese Art von Diagramm wurde gewählt, da sie zur Darstellung von Prozessen in der Softwareentwicklung geeignet ist.[53]

Es besteht aus verschiedenen Elementen, die bestimmte Funktionen repräsentieren. Ein Kreis markiert den Startpunkt, während ein Kreis innerhalb eines anderen Kreises den Endpunkt kennzeichnet.[54] Pfeile zeigen den Übergang von einem Punkt zum nächsten an.[55] Rechtecke mit abgerundeten Ecken stehen für Aktionen, die ausgeführt werden müssen.[56] Rauten symbolisieren Entscheidungspunkte, von denen verschiedene Abläufe abhängen, je nach Erfüllung bestimmter Bedingungen. Sie dienen auch als Zusammenführungspunkte mehrerer Abläufe.[57] Horizontale Balken stellen parallele Abläufe dar, von denen mehrere gleichzeitig ausgeführt werden können. Diese können wiederum von einem horizontalen Balken zusammengeführt werden.[58]

[53] Vgl. Kleuker, S. (2018). *Grundkurs Software-Engineering mit UML.* Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-19969-2. S.11.

[54] Vgl. ebd.

[55] Vgl. ebd.

[56] Vgl. ebd.

[57] Vgl. ebd.

[58] Vgl. ebd. S12.

Das Diagramm ist in verschiedene Bereiche unterteilt, die durch gestrichelte Linien abgegrenzt sind. Diese Bereiche repräsentieren die verschiedenen Zuständigkeiten innerhalb des Programms, welche hier durch dessen Module wahrgenommen werden.[59] Diese Module sind Teile des Programms, die spezifische Aufgaben zur Erfüllung der Gesamtfunktion übernehmen. Die Namen der entsprechenden Module sind vertikal am Rand des Diagramms angeordnet. Durch diese strukturierte Darstellung wird der Konvertierungsprozess übersichtlich dargestellt und ermöglicht eine klare Nachverfolgung der einzelnen Schritte und Zuständigkeiten.

[59] Vgl. ebd. S17.

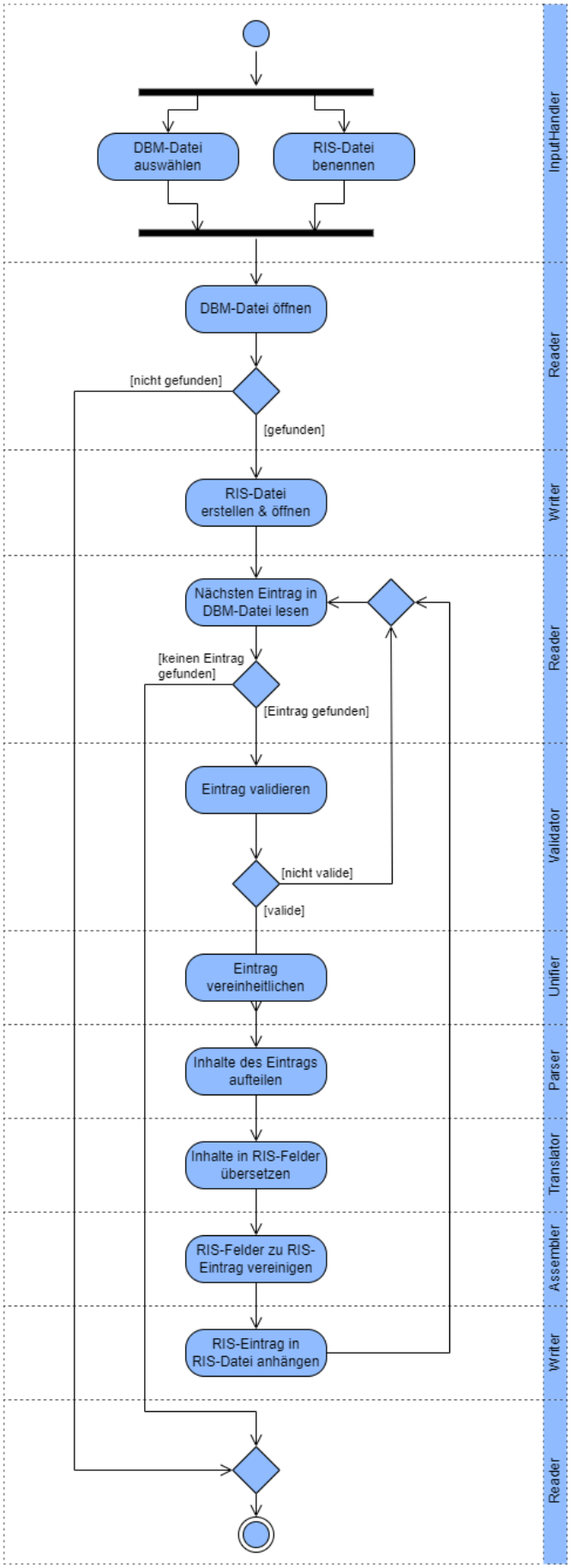

Abbildung 1: Aktivitätsdiagramm des Konvertierungsprogramms

Das InputHandler-Modul wählt die DBM-Datei aus, die konvertiert werden soll und benennt die Ziel-RIS-Datei. Da das Programm für die Nutzung über die Kommandozeile konzipiert ist, werden die Eingaben über Kommandozeilenargumente bereitgestellt, die beim Start des Programms übergeben werden. Es ist wichtig zu beachten, dass das Programm abbricht, wenn entweder keine oder mehr als zwei Kommandozeilenargumente übergeben werden.

Wenn genau zwei Kommandozeilenargumente übergeben werden und das erste Argument eine DBM-Datei und das zweite Argument eine RIS-Datei bezeichnet, leitet das InputHandler-Modul beide an die entsprechenden Module weiter, die für die weitere Verarbeitung zuständig sind. Andernfalls wird das Programm beendet.

Im Falle der Übergabe eines einzigen Kommandozeilenarguments, das eine DBM-Datei bezeichnet, wird automatisch eine RIS-Datei mit demselben Namen erstellt und beide werden anschließend an die entsprechenden Module übergeben. Wenn das Argument nicht korrekt ist wird das Programm ebenfalls beendet. Dieser Prozess stellt sicher, dass die Konvertierung nur dann erfolgt, wenn die erforderlichen Eingaben korrekt sind und verhindert potenzielle Fehler während der Ausführung.

Nachdem die DBM-Datei erfolgreich durch das InputHandler-Modul ausgewählt wurde, übernimmt das Reader-Modul die Aufgabe diese Datei zu öffnen und die darin enthaltenen Daten somit für die Extrahierung vorzubereiten. Da die DBM-Datei gemäß ISO 8859-1 kodiert ist[60], wird sie standardmäßig mit dieser Kodierung gelesen. Sollte es nicht möglich sein die angegebene DBM-Datei zu finden oder zu öffnen, wird das Programm beendet, um potenzielle Fehler zu vermeiden. Dieser Schritt gewährleistet, dass die Konvertierung reibungslos verläuft und die Daten ordnungsgemäß weiterverarbeitet werden können.

Nachdem die DBM-Datei erfolgreich geöffnet wurde, übernimmt das Writer-Modul die Aufgabe, die entsprechende RIS-Datei zu erstellen und zu öffnen, in welche die später konvertierten Daten geschrieben werden sollen. Die RIS-Datei wird standardmäßig mit der UTF-8-Kodierung erstellt, um eine breitere Unterstützung verschiedener Zeichen und Symbole sicherzustellen. Wenn die angegebene RIS-Datei bereits existiert wird sie überschrieben, um sicherzustellen, dass die neuen Daten korrekt eingefügt werden und keine Konflikte mit vorhandenen Dateien auftreten. Dieser Schritt ist entscheidend, um zu gewährleisten, dass die konvertierten Daten korrekt gespeichert werden und keine Datenverluste auftreten.

Nachdem die RIS-Datei erfolgreich erstellt wurde setzt das Programm den Prozess fort, indem das Reader-Modul den nächsten Eintrag im modifizierten Allegro-Neutralformat aus der DBM-Datei liest. Dieser Eintrag wird anschließend an das nächste Modul übergeben, das für die weitere Verarbeitung zuständig ist. Sollte jedoch kein Eintrag im Allegro-Neutralformat in der DBM-Datei gefunden werden können, beendet das

[60] Dies geht aus der Auflösung aller Kodierungsfehler innerhalb der Datei bei der Öffnung in dem entsprechenden Format hervor

Programm seine Ausführung. Diese Überprüfung stellt sicher, dass nur gültige Daten verarbeitet werden und dass der Konvertierungsprozess ordnungsgemäß beendet wird.

Nachdem ein Eintrag aus der DBM-Datei gelesen wurde, wird er dem Validator-Modul übergeben, um seine Gültigkeit zu überprüfen. Das Validator-Modul untersucht den Eintrag nach dem Feld "030". Sollte dieses Feld nicht vorhanden sein oder eine Verweisung enthalten, liest das Reader-Modul den nächsten Eintrag aus der DBM-Datei und der Übersetzungsprozess beginnt von vorn.

Wenn das Feld "030" vorhanden ist und keine Verweisung enthält, wird der Eintrag als gültig betrachtet und an das nächste Modul übergeben, das für die weitere Verarbeitung zuständig ist. Diese Überprüfung stellt sicher, dass nur Einträge mit den erforderlichen Informationen weiterverarbeitet werden und potenzielle Fehler vermieden werden. Verweisungen werden entsprechend der Datenbankuntersuchung in Kapitel "4 Untersuchung der Datenbank" ermittelt.

Nachdem der Eintrag erfolgreich validiert wurde, wird er dem Unifier-Modul übergeben, das den Eintrag zwecks einfacherer Verarbeitung vereinheitlicht und anschließend an das nächste Modul weitergibt. Dabei werden mehrere Schritte zur Durchführung dieses Prozesses unternommen:

- Die Feldbezeichnung und deren Doppelpunkt werden so umgewandelt, dass sie immer mit einem Leerzeichen von dem Feldinhalt getrennt sind, um dem Schema zu entsprechen.
- Alle Trennzeichen werden so verändert, dass Leerzeichen um sie herum entfernt werden.
- Alle Nicht-Sortierzeichen werden entfernt, um die Konsistenz zu gewährleisten.
- Im Feld "100" werden Doppelpunkte so umgewandelt, dass sie jeweils von einem Leerzeichen umgeben sind, um dem vorgegebenen Schema zu entsprechen, außer bei URLs, die anhand der Zeichenketten "http:/" oder "https://" erkannt werden können.
- Im Feld "411" werden Semikola so umgewandelt, dass sie jeweils von einem Leerzeichen umgeben sind, um dem vorgegebenen Schema zu entsprechen.
- Im Feld "700" werden Seitenangaben entfernt und in das Feld "810" überführt, um die Strukturierung der Daten zu optimieren.

Diese Anpassungen stellen sicher, dass die Daten einheitlich und gemäß den Anforderungen des Kategorienschemas strukturiert sind, was eine reibungslose weitere Verarbeitung ermöglicht.

Nachdem der Eintrag erfolgreich vereinheitlicht wurde, wird er dem Parser-Modul übergeben, das die Inhalte des Eintrags analysiert und entsprechend aufteilt, bevor es das Ergebnis an das nächste Modul weitergibt. Dabei wird ein Dictionary als Grundlage für die Aufteilung des Eintrags erstellt. Ein Dictionary ist eine Datenstruktur in der Programmierung, die einzelne Informationen Schlüsselwörtern als Identifikatoren

zuordnet.[61] Alle Felder des Eintrags werden gemäß den Ergebnissen aus Kapitel "3.3. Felduntersuchung" aufgeteilt. Wenn keine Aufteilung erforderlich ist, wird der gesamte Inhalt dem entsprechenden Feldnamen als Schlüsselwort im Dictionary zugeordnet. Wenn jedoch eine Aufteilung erforderlich ist, werden die einzelnen Bestandteile als Liste dem entsprechenden Feldnamen als Schlüsselwort im Dictionary zugeordnet. Dieser Prozess ermöglicht eine strukturierte Aufbereitung der Eintragsdaten, wodurch sie für weitere Verarbeitungsschritte besser zugänglich und analysierbar sind.

Nachdem das Dictionary erfolgreich erstellt wurde, wird es dem Translator-Modul übergeben, das den Inhalt in das RIS-Format übersetzt und anschließend an das nächste Modul weitergibt. Dabei wird ein neues Dictionary als Grundlage für die Übersetzung erstellt. Das Translator-Modul iteriert durch die Inhalte des übergebenen Dictionaries und ordnet sie entsprechend den Konkordanzen und den Ergebnissen aus Kapitel "3.3 Felduntersuchung" RIS-Feldern als Schlüsselwörter im neuen Dictionary zu. Da im RIS-Format Felder mehrmals genutzt werden können, werden die zugeordneten Inhalte als Liste im neuen Dictionary gespeichert. Dadurch können Inhalte bei einer erneuten Nutzung des Feldes hinzugefügt werden. Dieser Prozess gewährleistet eine korrekte Übersetzung der Eintragsdaten in das RIS-Format und stellt sicher, dass alle relevanten Informationen ordnungsgemäß strukturiert sind, um eine reibungslose Weiterverarbeitung zu ermöglichen.

Nachdem das Dictionary erfolgreich erstellt und übersetzt wurde, wird es dem Assembler-Modul übergeben, das aus dem Dictionary einen RIS-Eintrag zusammensetzt und diesen anschließend an das nächste Modul weitergibt. Das Assembler-Modul iteriert durch alle Schlüsselwörter des Dictionaries und erstellt für jeden Listeninhalt der zugeordneten Liste ein Feld, wobei das Schlüsselwort als Feldbenennung und der Listeninhalt als dessen Inhalt verwendet wird. Dabei wird immer mit dem Schlüsselwort "TY" begonnen, da RIS-Einträge mit diesem Feld beginnen müssen. Nachdem alle Schlüsselwörter durchlaufen wurden, wird der Eintrag durch das Feld "ER" abgeschlossen. Dieser Prozess stellt sicher, dass das RIS-Format korrekt und vollständig erstellt wird, sodass die Einträge gemäß den RIS-Spezifikationen strukturiert sind und problemlos von anderen Systemen weiterverarbeitet werden können.

Nachdem der RIS-Eintrag erfolgreich zusammengesetzt wurde, wird er dem Writer-Modul übergeben, das den Eintrag der RIS-Datei hinzufügt und abschließend eine leere Zeile einfügt, um die Trennung zwischen den Einträgen zu gewährleisten. Nach diesem Schritt liest das Reader-Modul den nächsten Eintrag aus der DBM-Datei, um den Konvertierungsprozess fortzusetzen. Dieser Zyklus wird wiederholt, bis alle Einträge aus der DBM-Datei erfolgreich in die RIS-Datei übertragen wurden. Dieser kontinuierliche Ablauf ermöglicht eine effiziente und zuverlässige Konvertierung der Datenbankinhalte in das gewünschte Format.

[61] Vgl. Saake, G., & Sattler, K.-U. (2021). *Algorithmen und Datenstrukturen: Eine Einführung mit Java* (6. überarbeitete und erweiterte Auflage). dpunkt.verlag. S.361.

5.2 Implementierung

Das Programm wurde für diese Arbeit in der Programmiersprache Java implementiert und als Maven-Projekt verwaltet. Die Struktur und Funktionsweise des Projekts entsprechen der Beschreibung im letzten Kapitel. Dieser Ansatz ermöglicht eine effiziente Verwaltung der Abhängigkeiten und eine standardisierte Projektstruktur, was zur einfacheren Entwicklung, Wartung und Skalierbarkeit beiträgt. Die Verwendung von Maven erleichtert auch die Integration von Bibliotheken und externen Ressourcen, was die Entwicklung des Programms erleichtert und die Arbeitsabläufe optimiert.[62]

Der Source-Code des Programms ist in den beigefügten Anlagen unter dem Dateipfad "converter/src/main/java" zu finden. Um das Programm zu verwenden, wird eine Java-Laufzeitumgebung der Version 17 oder höher benötigt.[63] Das Programm trägt den Namen "converter-1.0.0.jar" und befindet sich im Dateipfad "converter/target". Es ist jedoch möglich, die Platzierung des Programms nach Belieben zu ändern.

Um das Programm auszuführen, kann folgender Befehl in der Kommandozeile verwendet werden: "<Pfad/>java -jar converter-1.0.0.jar <dbm> [<ris>]".

Hierbei muss "<Pfad\>" durch den aktuellen Dateipfad zum Programm ersetzt werden. "<dbm>" sollte durch den aktuellen Dateipfad zu der zu übersetzenden DBM-Datei ersetzt werden. Optional kann "[<ris>]" weggelassen oder durch den aktuellen Dateipfad mit dem Namen einer RIS-Datei ersetzt werden, an dem die zu erzeugende RIS-Datei mit dem gewünschten Namen erstellt werden soll.

Das Programm ist mit der entsprechenden Laufzeitumgebung bereits ausführbar, kann aber auch auf zwei verschiedene Arten selber kompiliert werden:

1. Wenn Apache Maven mit Version 3.9.6 oder höher installiert ist, kann die Kompilierung mit folgendem Befehl in der Kommandozeile ausgeführt werden: "<Pfad/>mvn install".

   Apache Maven steht unter der Apache-Lizenz 2.0, die es erlaubt, diese Software frei in jedem Umfeld zu verwenden, zu modifizieren und zu verteilen.[64] "<Pfad\>" muss durch den aktuellen Dateipfad zum Hauptverzeichnis des Repositoriums ersetzt werden, wobei wichtig ist, dass sich in diesem die Datei "pom.xml" befindet.

2. Wenn Apache Maven mit Version 3.9.6 oder höher nicht installiert ist, kann die Kompilierung stattdessen durch folgenden Befehl in der Kommandozeile ausgeführt werden: "javac -encoding UTF-8 <Pfad/>Main.java".

---

[62] Vgl. The Apache Software Foundation. (o.D.). *Maven: Introduction.* Maven. https://maven.apache.org/what-is-maven.html. [Letzter Aufruf: 11.03.2024].

[63] Vgl. Oracle. (25.01.2022). *JDK 17 documentation: home.* Oracle Help Center. https://docs.oracle.com/en/java/javase/17/. [Letzter Aufruf: 11.03.2024].

[64] Vgl. The Apache Software Foundation. (o.D.). *Apache License, Version 2.0.* Apache. https://www.apache.org/licenses/LICENSE-2.0. [Letzter Aufruf: 11.03.2024].

"<Pfad\>" muss durch den aktuellen Dateipfad zum Hauptverzeichnis des Source-Codes ersetzt werden, wobei wichtig ist, dass sich in diesem die Datei "Main.java" befindet. Dadurch entstehen im Verzeichnis des Source-Codes ausführbare Dateien, die durch folgenden Befehl in der Kommandozeile ausgeführt werden können: "java <Pfad/>Main <dbm> [<ris>]".

Hierbei muss "<Pfad\>" durch den aktuellen Dateipfad zum Hauptverzeichnis des Source-Codes ersetzt werden, wobei wichtig ist, dass sich in diesem die Datei "Main.class" befindet.

Zusätzlich existiert innerhalb des Dateipfades "converter/src/test/java" in den beigefügten Anlagen Test-Code, der dazu dient, die Funktionalität des Programms sicherzustellen. Dabei werden die einzelnen Module des Programms isoliert und unter verschiedenen Szenarien getestet. Diese Tests werden automatisch durch Apache Maven während der Kompilierung durchgeführt. Um die Tests manuell auszuführen, kann der folgende Befehl in der Kommandozeile verwendet werden: "<Pfad/>mvn test".

Hierbei muss "<Pfad\>" durch den aktuellen Dateipfad zum Hauptverzeichnis des Test-Codes ersetzt werden. Wenn während der Ausführung keine Fehlermeldungen erscheinen, waren die Tests erfolgreich. Dieser Testprozess trägt dazu bei die Qualität des Programms sicherzustellen und potenzielle Fehler frühzeitig zu erkennen und zu beheben.

Die durch das implementierte Programm konvertierte Datenbank in das RIS-Format kann den beigefügten Anlagen im Ordner "data" als "litie34.ris" entnommen werden. Diese Datei enthält die Datenbankinformationen im RIS-Format, die durch das Programm konvertiert wurden. Sie kann für weitere Analysen, Importe oder andere Zwecke verwendet werden, die mit dem RIS-Format kompatibel sind.

6 Ergebnis der Konvertierung

In diesem Kapitel soll das Ergebnis der Konvertierung durch das zuvor erstellte Konvertierungsprogramm untersucht werden. Dazu werden Einträge der ursprünglichen Datenbank im modifizierten Allegro-Format mit Einträgen der konvertierten Datenbank im RIS-Neutralformat verglichen. Dieser Vergleich dient als Erfolgskontrolle und bildet eine Grundlage für die Beantwortung der eingangs gestellten Forschungsfrage. Durch die Analyse der konvertierten Datenbank können etwaige Unterschiede, Verluste oder Veränderungen identifiziert werden, um die Zuverlässigkeit und Genauigkeit des Konvertierungsprozesses zu bewerten und Rückschlüsse auf die Qualität der Konvertierung zu ziehen.

Da nicht alle Einträge der Datenbank, wie bereits in der Einleitung erwähnt, untersucht werden können, wird die Untersuchung anhand einer Stichprobe durchgeführt. Dabei wird eine proportionale Schichtenstichprobe verwendet. Dies ist eine Stichprobenmethode, bei der eine Grundgesamtheit in verschiedene Schichten unterteilt wird, um repräsentative Stichproben aus jeder Schicht zu ziehen.[65] Im vorliegenden Kontext bedeutet "proportional", dass die Größe der Schichten in der Stichprobe proportional zu der Größe der Schichten in der Grundgesamtheit ist.[66]

Einträge werden entsprechend ihrer Schicht zufällig aus der Datenbank ausgewählt. Das relevante Merkmal bei der Schichtung der Stichprobe ist der Dokumenttyp der Einträge, um möglichst viele verschiedene Feldbelegungen und seltene Fälle zu berücksichtigen. Es wird immer nur der erstgenannte Dokumenttyp betrachtet, da andere Dokumenttypen bei mehreren Inhalten im Konvertierungsprozess verworfen wurden.

Einträge mit Verweisungen als Dokumenttyp oder ohne Dokumenttyp wurden nicht in das RIS-Format übersetzt und sind daher nicht Teil der Grundgesamtheit. Die Grundgesamtheit beträgt somit 43.412 Einträge.[67] Dieser Ansatz ermöglicht eine repräsentative Auswahl von Einträgen für die Untersuchung, wodurch die Ergebnisse auf die gesamte Datenbankpopulation verallgemeinert werden können.

Der Anteil der Dokumenttypen in der Datenbank "Literatur zur Informationserschließung" kann der folgenden Tabelle entnommen werden, bei der die prozentualen Anteile auf zwei Nachkommastellen gerundet wurden:[68]

Dokumenttyp	**Anteil (absolut)**	**Anteil (prozentual)**
a (Artikel/Aufsatz)	34.545	79,57%
ag (Artikel)	2	0,00%

[65] Vgl. Gillhofer, M. M. (2010). *Teilnehmer-Rekrutierung in der Online-Sozialforschung.* Eul. S68f.

[66] Vgl. ebd. S.69.

[67] Diese Menge kann durch den regulären Ausdruck "^030:.+$" minus der Größe des Ergebnisses durch den regulären Ausdruck "^030:.*(?:s\.|siehe).*$" ermittelt werden

[68] Diese Anteile können mit dem regulären Ausdruck "^030:x(?:\|[\w?]+)?(?:\|[\w?]+)?(?:\|[\w?]+)?$" ermittelt werden, wobei "x" in dem regulären Ausdruck durch den entsprechenden Dokumenttyp ausgetauscht werden muss

Dokumenttyp	Anteil (absolut)	Anteil (prozentual)
au (Konferenzschrift)	1	0,00%
b (Bibliographie)	52	0,12%
d (Dissertation)	35	0,08%
el (Elektronisches Dokument)	1.479	3,40%
fi (Mikroform)	11	0,02%
h (Anleitung)	53	0,12%
i (Nachschlagewerk/Informationsmittel)	132	0,30%
l (Loseblattsammlung)	39	0,09%
m (Monographie)	4423	10,19%
ms (Teil einer monographischen Reihe)	1	0,00%
n (Norm)	104	0,24%
p (Neudruck)	88	0,20%
pat (Patent)	4	0,00%
r (Bericht)	324	0,74%
s (Sammelwerk)	1206	2,78%
u (Skript/Unterrichtsmaterialien)	22	0,05%
vi (Video/Film)	1	0,00%
x (Hausarbeit/Diplomarbeit)	698	1,61%
z (Zeitschrift)	9	0,02%
? (Unbekannt)	180	0,41%

Tabelle 7: Anteile der Dokumenttypen in der Datenbank "Literatur zur Informationserschließung"

Die Größe der Schichten entspricht den jeweiligen ganzzahligen Prozentanteilen der Dokumenttypen in der Datenbank. Dokumenttypen, die einen prozentualen Anteil von unter Eins haben, werden auf Eins aufgerundet, um sie in die Stichprobe aufnehmen zu können, wodurch sie eine Größe von 112 Elementen annimmt. Aufgrund der Schichtung wird die Stichprobe mit dieser Größe als repräsentativ bezüglich der Datenbank betrachtet.[69]

Alle Ergebnismengen der Anteilbestimmung der Dokumenttypen werden jeweils in ein Textdokument kopiert. Mithilfe eines Zufallsgenerators werden dann zufällige Zeilen aus diesem Dokument ausgewählt. Da Notepad++ die Ergebnisse immer zusammen mit ihren Zeilennummern angibt, kann aus jedem Element der Ergebnismenge auf dessen Eintrag geschlossen werden.[70] Auf diese Weise wird die Stichprobe realisiert. Durch das

[69] Vgl. Statista. (o.D.). *Repräsentativität: Statista Definition.* Statista Lexikon. https://de.statista.com/statistik/lexikon/definition/116/repraesentativitaet/. [Letzter Aufruf: 11.03.2024].

[70] Vgl. Notepad++. (o.D.). *Searching: Notepad++ User Manual.*

Zufälligkeitsprinzip wird sichergestellt, dass die ausgewählten Einträge repräsentativ für die Gesamtpopulation sind und somit die Validität der Untersuchungsergebnisse gewährleistet wird.

Die Einträge der Stichprobe werden daraufhin geprüft, ob der richtige Dokumenttyp verwendet wurde, die richtigen Felder genutzt wurden und die Feldinhalte korrekt aufgeteilt wurden. Dieser Prozess prüft, ob die konvertierten Einträge im RIS-Format den erforderlichen Standards entsprechen und keine Informationen verloren gegangen sind.

Die Ergebnisse der Stichprobenuntersuchung können folgender Tabelle entnommen werden:[71]

Prüfung	**Fehlerhafte Einträge (absolut)**	**Fehlerhafte Einträge (prozentual)**
Dokumenttyp	0	0%
Feldnutzung	1	0,89%
Feldaufteilung	7	6,25%

Tabelle 8: Ergebnisse der Stichprobenprüfung des Konvertierungsprozesses

Alle Einträge wurden in die richtigen Dokumenttypen konvertiert, was auf eine erfolgreiche Umsetzung des Konvertierungsprozesses hinweist. In einem Eintrag wurde jedoch ein Feld nicht korrekt konvertiert, was 0,89% der Stichprobe entspricht. Dabei handelt es sich um ein Vorkommen des Feldes "660" in der ursprünglichen Datenbank. Dieses Feld ist nicht Teil der Dokumentation und wurde auch bei der Datenbankuntersuchung nicht gefunden. Aus diesem Grund konnte es nicht von dem Konvertierungsprogramm berücksichtigt werden. Es ist wichtig zu beachten, dass dieser Fehler eine geringe Prozentzahl der Stichprobe betrifft und die Gesamtheit der konvertierten Einträge weiterhin die richtigen Felder benutzen.

In sieben Einträgen kam es zu Fehlern bei der Aufteilung von Feldinhalten, was 6,25% der Stichprobe entspricht:

1. Bei den Datumsangaben "2003, Fall" und "Winter 1979/80" im Feld 700 kam es zu keiner Identifikation und somit keiner Aufteilung. Das Programm erfordert Datumsangaben im Format TT.MM.JJJJ, wobei Platzhalter und Datumsspannen möglich sind. Der teilweise natürlichsprachige Aufbau und das Fehlen von Tages- und Monatsangaben verhindern das Erkennen durch das Programm. Diese Restriktion ist notwendig, um Datumsangaben von anderen Inhalten des Feldes "700" zu unterscheiden.
2. Die Bandangaben "WB1" und "[=Suppl.20]" im Feld "700" wurden ebenfalls nicht erkannt, was zu keiner Feldaufteilung führte. Der unregelmäßige Aufbau der Bandangaben verhindert die Erkennung durch das Programm.

[71] Diese kann den beigefügten Anlagen in dem Ordner "test" als "stichprobe.txt" entnommen werden, in der die zu vergleichenden Einträge immer untereinandergeschrieben sind

3. Der Inhalt "Bergische Landeszeitung. Nr.1 vom 2.1.1998, S" im Feld "700" wurde nicht aufgeteilt, möglicherweise aufgrund einer unvollständigen Seitenangabe, die nicht von dem Programm erkannt werden konnte.
4. Der Inhalt einer URL im Feld "875" wurde korrekterweise als URL erkannt und ins Feld "UR" übersetzt. Allerdings wurde der Satz nicht von der URL getrennt, da URLs lediglich an der Zeichenkette "http://" oder "https://" erkannt werden und alle Zeichen dahinter als Teil der URL betrachtet werden.
5. Der Inhalt "Aslib proceedings. (bis 2012). Fortgesetzt als: Aslib journal of information management. 65(2013), ff" im Feld "700" wurde ebenfalls nicht aufgeteilt, aufgrund des deutlich unterschiedlichen Aufbaus im Vergleich zu anderen Feldern mit der Feldbenennung "700".

Durch das Ergebnis der Stichprobenuntersuchung kann davon ausgegangen werden, dass 7,14% der konvertierten Einträge der Datenbank einen Fehler enthalten, der in den meisten Fällen auf eine fehlerhafte Aufteilung eines Feldes zurückzuführen ist.

Da die in der Stichprobe gefundenen Fehler Sonderfälle darstellen, kann angenommen werden, dass viele weitere Sonderfälle existieren und ihre Korrektur somit vermutlich keine allgemeine Verbesserung des Konvertierungsprozesses bedeuten würde. Dies verdeutlicht jedoch die Grenzen einer automatisierten Lösung, die zwar die Bewältigung der Konvertierung der Datenbank effizient ermöglicht, jedoch nicht so effektiv ist wie eine manuelle Konvertierung. Dies liegt daran, dass durch die Automatisierung bekannte und nicht miteinander in Konflikt stehende Regelmäßigkeiten vorausgesetzt werden, wie die Untersuchung der Datenbank bereits gezeigt hatte.

Weil die gefundenen Fehler nicht zu einer Verwerfung des Inhalts, sondern lediglich zu einer falschen Nutzung von RIS-Feldern geführt haben, führen diese zu keinem Informationsverlust, sondern nur zu einem Qualitätsverlust. Dies unterstreicht die Bedeutung einer sorgfältigen Überprüfung und gegebenenfalls manuellen Nachbearbeitung der konvertierten Datenbank, um deren Qualität zu gewährleisten.

7 Fazit

Nicht alle Bestandteile des modifizierten Allegro-Neutralformats können direkt in das RIS-Format übersetzt werden. Es muss ein gewisser Informationsverlust akzeptiert werden, der daraus resultiert, dass das RIS-Format bibliografische Objekte anders darstellt als das modifizierte Allegro-Neutralformat. Dies liegt daran, dass das modifizierte Allegro-Neutralformat ein bibliothekarisches Datenformat ist, während das RIS-Format ein Referenzdatenformat darstellt, welche entsprechend unterschiedliche Zwecke verfolgen.

Dieser Unterschied macht sich besonders bei den Feldern für bibliothekarische Sachschlagworte, Klassifikationen und Notationen im modifizierten Allegro-Neutralformat bemerkbar. Im RIS-Format hingegen ist die Abhängigkeit der meisten Felder von ihrem jeweiligen Dokumenttyp und das Fehlen der zuvor genannten bibliothekarischen Kategorien bemerkbar. Diese unterschiedlichen Strukturen und Ziele führen zu einer Herausforderung bei der direkten Übersetzung zwischen den Formaten und erfordern eine sorgfältige Berücksichtigung der Datenstruktur und des Informationsgehalts für den Konvertierungsprozess.

Aufgrund dieser Situation sind Entscheidungen bezüglich des Umgangs mit dem Informationsverlust erforderlich gewesen, was innerhalb dieser Arbeit bedeutete, dass versucht wurde, möglichst viele Informationen zu erhalten. Diese Entscheidung basiert auf dem Ziel zum langfristigen Erhalt der Datenbank beizutragen. Daher wurden viele Inhalte der Datenbank in das RIS-Feld für Notizen übersetzt, wenn diese ansonsten keine geeigneten Darstellungsmöglichkeiten im RIS-Format hatten. Ebenso wurde auf die eigentlich erforderliche Aufteilung des Inhalts eines Feldes des modifizierten Allegro-Neutralformats auf mehrere Felder des RIS-Formats verzichtet, wenn eine zuverlässige Aufteilung nicht gewährleistet werden konnte.

Dennoch muss anerkannt werden, dass Informationsverlust unvermeidlich ist, insbesondere, wenn Dokumenttypen und Schlagwortkategorien zu allgemeineren Kategorien zusammengefasst werden müssen. Diese Entscheidung zugunsten der Vollständigkeit der Datenbank geht jedoch zu Lasten der Qualität der RIS-Daten. Es ist wichtig, diesen Kompromiss zwischen Vollständigkeit und Qualität bei der Konvertierung der Datenbank zu berücksichtigen und die Auswirkungen auf die Endnutzung der Datenbank sorgfältig abzuwägen.

Auf Grundlage von Regelmäßigkeiten im Aufbau von Einträgen der Datenbank, wie der Kombination von Feldern, Bestandteilen von Feldinhalten und deren Trennung, kann eine automatisierte Verarbeitung und Übersetzung erfolgen. Problematisch ist jedoch die Identifikation dieser Regelmäßigkeiten, da sie eine umfangreiche Untersuchung der Datenbank und ihres Aufbaus erfordert. Dies wird umso problematischer, je größer die Datenbank ist und je heterogener ihr Aufbau ausfällt.

Es müssen verschiedene mögliche Muster für denselben Sachverhalt ermittelt und berücksichtigt werden, wobei darauf geachtet werden muss, dass sie nicht miteinander

in Konflikt stehen. Dies kann nicht immer sichergestellt werden. In solchen Fällen muss erneut eine Entscheidung bezüglich eines möglichen Informationsverlustes getroffen werden, so etwa bei dem Verzicht auf Feldaufteilungen, wie bereits erwähnt.

Die Untersuchung der Stichproben auf die Qualität der Konvertierung ergibt einen klaren Erfolg der erarbeiteten Lösungen, weist jedoch auch Fehler bei der Konvertierung mancher Einträge auf, die auf Unregelmäßigkeiten im Aufbau von Feldinhalten innerhalb der Datenbank zurückzuführen sind. Diese Unregelmäßigkeiten konnten bei der Untersuchung der Datenbank nach Regelmäßigkeiten nicht aufgefunden werden, da sie Sonderfälle darstellen und daher in der Menge der Einträge untergehen. Sie entziehen sich automatisierten Lösungen, da sie nicht den grundlegenden Mustern der Datenbank entsprechen. Die konkrete Berücksichtigung dieser Sonderfälle hat nur einen geringen Einfluss auf die Qualität der Gesamtkonvertierung. Die Iteration von Stichprobenverfahren zur Auffindung und Verbesserung weiterer Sonderfälle kann die Qualität der Datenbank weiter verbessern, ist jedoch mit einem hohen zeitlichen Aufwand verbunden.

Dies verdeutlicht die Grenzen einer automatisierten Lösung, die zwar die Bewältigung der Konvertierung der Datenbank effizient ermöglicht, jedoch nicht so effektiv ist wie eine manuelle Konvertierung, da sie Regelmäßigkeiten voraussetzt, die bekannt sein müssen und miteinander nicht in Konflikt stehen dürfen. Dies bedeutet einen weiteren Informationsverlust, der akzeptiert werden muss.

Zusammenfassend lässt sich sagen, dass die Konvertierung bibliographischer Referenzdaten in ein neutrales Austauschformat erfolgreich durch automatisierte Lösungen möglich ist. Jedoch sind Unregelmäßigkeiten im Aufbau der Datenbank und Unterschiede zwischen dem Ausgangs- und Zielformat entscheidend für mögliche Informationsverluste. Die automatisierte Übersetzung erfordert daher eine genaue Untersuchung der Datenbankstruktur und kann dennoch nicht alle Sonderfälle erfassen. Die Qualität der Konvertierung kann durch iterative Stichprobenverfahren verbessert werden, jedoch ist dies mit einem hohen zeitlichen Aufwand verbunden. Letztendlich müssen Entscheidungen darüber getroffen werden, wie mit dem auftretenden Informationsverlust umgegangen werden soll, wobei eine Abwägung zwischen Effizienz und Qualität erfolgen muss.

Literaturverzeichnis

[Aurimasv]. (12.04.2012). *Ris tag map: aurismasv/Translators Wiki.* GitHub. https://github.com/aurimasv/translators/wiki/RIS-Tag-Map. [Letzter Aufruf: 11.03.2024].

Beiler, C., Gratzl, P., Schubert, B., Steiner, C., & Steltzer, R. (24.11.2018). *Erschließungsarbeit in Alma: Erfahrungen aus dem OBV vor, während und nach der Aleph-Ablöse.* Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare, 71(2), 282–306. https://doi.org/10.31263/voebm.v71i2.2134.

Bertram, J. (2019). *Abschlussarbeiten in der Bibliotheks- und Informationswissenschaft.* De Gruyter Saur.

Gillhofer, M. M. (2010). *Teilnehmer-Rekrutierung in der Online-Sozialforschung.* Eul.

Gödert, W., Lepsky, K., & Nagelschmidt, M. (2012). *Informationserschließung und automatisches Indexieren: Ein Lehr- und Arbeitsbuch.* Springer.

Ho, D. (o.D.). *What is notepad++.* Notepad++. https://notepad-plus-plus.org/. [Letzter Aufruf: 11.03.2024].

Kann, B. (24.11.2018). *Alma im Österreichischen Bibliothekenverbund (OBV): Aus der Werkstatt der OBVSG.* Mitteilungen der Vereinigung Österreichischer Bibliothekarinnen und Bibliothekare, 71(2), 307–319. https://doi.org/10.31263/voebm.v71i2.2133.

Kleuker, S. (2018). *Grundkurs Software-Engineering mit UML.* Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-19969-2.

[larsgw], [dstillman], [adamsmith]. (12.04.2021-13.04.2021). *Ris specification.* Zotero Forums. https://forums.zotero.org/discussion/89035/ris-specification. [Letzter Aufruf: 11.03.2024].

Lepsky, K. (05.2017). *Kategorienschema der Datenbank "Literatur zur Informationserschließung".* https://ixtrieve.fh-koeln.de/lehre/kategorienschema-litac.pdf. [Letzter Aufruf: 11.03.2024].

Lepsky, K. (o.D.). *Literatur: Literaturdatenbank zu den Themen Informationserschließung und Information Retrieval.* Indexierung-Retrieval. https://www.indexierung-retrieval.de/2013/02/literatur.html. [Letzter Zugriff: 11.03.2024].

Library of Congress Network Development and MARC Standards Office. (o.D.). *MARC Specialized Tools.* MARC Records, Systems, and Tools (Network Development and MARC Standards Office, Library of Congress). https://www.loc.gov/marc/marctools.html. [Letzter Aufruf: 11.03.2024].

Microsoft. (29.11.2024). *Durchführen von Komponententests Mithilfe des test-explorers - visual studio (windows).* Microsoft Learn. https://learn.microsoft.com/de-de/visualstudio/test/unit-test-basics?view=vs-2022. [Letzter Aufruf: 11.030.2024].

Microsoft. (12.06.2023). *Sprachelemente für reguläre Ausdrücke – Kurzübersicht - .NET.* Microsoft Learn. https://learn.microsoft.com/de-de/dotnet/standard/base-types/regular-expression-language-quick-reference. [Letzter Aufruf: 11.03.2024].

Neumann, E. (o.D.). *allegro-C: Das Neutralformat.* allegro. https://www.allegro-c-support.de/doku/neutral/. [Letzter Aufruf: 11.03.2024].

Neumann, E. (o.D.). *allegro-C: Das Neutralformat: Ausführliche Liste aller Felder.* allegro. https://www.allegro-c-support.de/doku/neutral/tab.htm. [Letzter Aufruf: 11.03.2024].

Notepad++. (o.D.). *Searching: Notepad++ User Manual.* Notepad++ user manual. https://npp-user-manual.org/docs/searching/. [Letzter Aufruf: 11.03.2024].

Oracle. (25.01.2022). *JDK 17 documentation: home.* Oracle Help Center. https://docs.oracle.com/en/java/javase/17/. [Letzter Aufruf: 11.03.2024].

Oracle. (o.D.). *What is Java and why do I need it?.* Java.com. https://www.java.com/en/download/help/whatis_java.html. [Letzter Aufruf: 311.03.2024].

Plaum, C. (08.11.2022). *Die Hochschulbibliotheken in NRW auf dem Weg in die Alma-Cloud.* ABI Technik, 42(4), 265–271. https://doi.org/10.1515/abitech-2022-0046.

Python. (10.2023). *Python release python 3.11.6.* https://www.python.org/downloads/release/python-3116/. [Letzter Aufruf: 11.03.2024].

Python. (o.D.). *Welcome to Python.org.* https://www.python.org/about/. [Letzter Aufruf: 11.03.2024].

Saake, G., & Sattler, K.-U. (2021). *Algorithmen und Datenstrukturen: Eine Einführung mit Java* (6. überarbeitete und erweiterte Auflage). dpunkt.verlag.

Statista. (o.D.). *Repräsentativität: Statista Definition.* Statista Lexikon. https://de.statista.com/statistik/lexikon/definition/116/repraesentativitaet/. [Letzter Aufruf: 11.03.2024].

The Apache Software Foundation. (o.D.). *Apache License, Version 2.0.* Apache. https://www.apache.org/licenses/LICENSE-2.0. [Letzter Aufruf: 11.03.2024].

The Apache Software Foundation. (o.D.). *Maven: Introduction.* Maven. https://maven.apache.org/what-is-maven.html. [Letzter Aufruf: 11.03.2024].

The Thomson Corporation. (06.10.2011). *RIS format specifications.* Wayback Machine. https://web.archive.org/web/20120526103719/http://refman.com/support/risformat_intro.asp. [Letzter Aufruf: 11.03.2024].

The Thomson Corporation. (07.07.2024). *RIS format specifications: Tag Definitions: Title and Reference Type Tags.* Wayback Machine. https://web.archive.org/web/20100726184137/http://www.refman.com/support/risformat_tags_01.asp. [Letzter Zugriff: 11.03.2024].

The Tomson Corporation. (14.02.2004). *RIS format specifications: Tag Format.* Wayback Machine. https://web.archive.org/web/20110926022719/http://www.refman.com/support/risformat_fields_01.asp. [Letzter Aufruf: 11.03.2024].

Wikipedia. (12.02.2024). *Ris (file format).* https://en.wikipedia.org/wiki/RIS_(file_format). [Letzter Aufruf: 11.03.2024].

Erklärung

Ich versichere, die von mir vorgelegte Arbeit selbstständig verfasst zu haben. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten oder nicht veröffentlichten Arbeiten Anderer oder des Verfassers selbst entnommen sind, habe ich als entnommen kenntlich gemacht. Sämtliche Quellen und Hilfsmittel, die ich für die Arbeit benutzt habe, sind angegeben. Die Arbeit hat mit gleichem Inhalt bzw. in wesentlichen Teilen noch keiner anderen Prüfungsbehörde vorgelegen.