Publisher: Disambiguierung und Historisierung
- Projektbericht -

Andre Bruns, Christopher Lenke,
Christine Rimmert

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1 Projektziel

Die Landschaft der Wissenschaftsverlage unterliegt seit mehr als drei Dekaden einer starken Veränderungsdynamik. Neben einer regelmäßig stattfindenden Akquise kleinerer und mittlerer Verlage durch größere Player ist insbesondere in der jüngeren Vergangenheit auch der Zusammenschluss von großen Anbietern zu beobachten. Das sicherlich bekannteste Beispiel der letzten Jahre ist die Fusionierung von MacMillan Science mit dem Verlag Science + Business Media. In den meisten Fällen erfolgte eine Übernahme aller vorhandenen Zeitschriften und Buchtitel in den Besitz und Bestand des aufkaufenden Verlages, so dass große Konzerne mit einer verzweigten Organisationsstruktur und einer Vielzahl von Imprints entstanden sind[1]. Diese Entwicklung hat auch Konsequenzen für die Struktur der Anbieterseite des Markts für wissenschaftliche Publikationen.

In den einschlägigen Bibliometriedatenbanken finden sich zwar Informationen zur früheren und gegenwärtigen Zugehörigkeit von Publikationen zu Wissenschaftsverlagen, allerdings lagen diese bislang nicht in bereinigter Form vor und auch die Beziehungen zwischen unterschiedlichen Verlagseinheiten bildete sich nicht in ihnen ab. Daher eigneten sich diese Datenbanken bisher kaum, um die angesprochenen Entwicklungen empirisch zu untersuchen. Ziel des Projekts „Publisher: Disambiguierung und Historisierung“ ist es daher, die im Web of Science (WoS) und Scopus vorliegenden Verlagsinformationen zu bereinigen und anzureichern, so dass sich Struktur und Entwicklung der heute größten Wissenschaftsverlage in ihnen abbilden. Die bereinigten Verlagsinformationen sollen dabei in entsprechenden Tabellen in den zur Projektlaufzeit aktuellen Bibliometriedatenbanken des Kompetenzzentrums Bibliometrie[2] bereitgestellt werden.

2 Datengrundlage

Für dieses Projekt wurden die beiden Datenbanken des Kompetenzzentrums Bibliometrie verwendet, die am FIZ Karlsruhe – Leibniz-Institut für Informationsinfrastruktur[3] gehostet werden. Dies sind:

- die Web of Science Core Collection von Thomson Reuters (heute: Clarivate Analytics). Diese beinhaltet die Core Collection, eine umfassende Sammlung, die den Zeitraum von 1980 bis zur 17. Kalenderwoche 2017abdeckt[4] und den ‚Science Citation Index Expanded’ (ab 1980), den ‚Social Sciences Citation Index Expanded’ (ab 1980), den ‚Arts & Humanities Citation Index’ (ab 1980) sowie die ‚ISI Pro-

[1] RELX Group, Springer Nature, John Wiley & Sons, Wolters Kluwer und Taylor & Francis

[2] Vgl. Kompetenzzentrum Bibliometrie

[3] Zugriff auf diese Daten besteht nur für Mitglieder des Kompetenzzentrum Bibliometrie. Siehe: http://www.bibliometrie.info/.

[4] Diese Einschränkung war notwendig, da die Bibliometriedatenbank des FIZ zu einem fixen Termin im März 2017 festgeschrieben wurde und keine Veränderungen für das restliche Jahr mehr mit aufgenommen wurden.

ceedings – Science and Technology' (ab 1990) und die ‚ISI Proceedings – Social Sciences and Humanities' (ab 1990),

- die Scopus Bibliometriedatenbank des Anbieters Elsevier, von 2004 bis Dezember 2017.

Die in beiden Datenbanken vorhandenen Verlagsbezeichnungen bilden zusammen mit den zugehörigen Adressen die Datengrundlage. Eine Verlagsbezeichnung wird hier nicht bereits als Name eines Verlages betrachtet, sondern zunächst als eine Zeichenfolge (String). Aufgrund der Vielzahl an Varianten und Änderungen der Verlagsbezeichnung im Laufe der Zeit sollen diese Strings im Projekt jeweils einer ‚Verlagsentität' (ein real existierender Verlag) mit einem standardisierten Verlagsnamen und einem Identifier zugeordnet werden.

Aus den im Web of Science (WoS) und in Scopus vorhandenen Verlagsbezeichnungen wurde jeweils eine Tabelle mit distinkten Verlagsbezeichnungen mit Auftrittshäufigkeiten und mini- sowie maximalem Publikationsjahr erstellt, die als Arbeitsgrundlage für die Disambiguierung diente. Adressinformationen wurden dabei im Verfahren hinzugezogen sofern dies erforderlich war. Die generierte Arbeitstabelle enthält 23.962 distinkte Verlagsbezeichnungen (ohne die Berücksichtigung von Adressen). Tabelle 1 gibt die Anzahl distinkter Verlagsbezeichnungen wieder und verweist auf eine geringe Schnittmenge der in beiden Datenbanken enthaltenen Verlagsbezeichnungen:

Datenbank	Anzahl distinkte Verlagsbezeichungen
WoS	11.815
Scopus	13.047
WoS ∩ Scopus	900

Tabelle 1: Distinkte Verlagsbezeichnungen für WoS und Scopus

Der Grund dafür liegt nicht etwa darin, dass nur wenige Verlage in beiden Datenbanken vorkommen, sondern vielmehr in Unterschieden in der Bezeichnung desselben Verlags in den beiden Datenquellen. Zurückzuführen ist dies vor allem auf die Vorstandardisierung der Daten, die das Web of Science vornimmt. Das folgende Beispiel (Abb.1) zeigt Verlagsbezeichnungen, die den Substring 'WILEY' enthalten, mit ihren Auftrittshäufigkeiten im WoS und Scopus. Die Verlagsbezeichnungen und deren Häufigkeiten zwischen den Datenbanken unterscheiden sich deutlich. Insbesondere treten viele Verlagsbezeichnungen aus Scopus im WoS gar nicht auf. Erkennbar sind dabei auch Beispiele für die Auswirkungen der Vorstandardisierung im WoS: während 'JOHN WILEY & SONS LTD' im WoS vorkommt, nicht aber 'JOHN WILEY AND SONS LTD' (Vorstandardisierung: 'AND' wird zu &), existieren in Scopus beide Varianten.

Verlagsbezeichnung	#WoS	#Scopus
WILEY-BLACKWELL	1558758	34278
WILEY-LISS	349731	0
WILEY-V C H VERLAG GMBH	291210	0
JOHN WILEY & SONS LTD	226252	4172
JOHN WILEY & SONS INC	221271	9020
WILEY-BLACKWELL PUBLISHING, INC	184900	0
WILEY	29988	15703
WILEY PERIODICALS, INC	15383	0
SCRIPTA TECHNICA-JOHN WILEY & SONS	8146	0
WILEY-LISS, INC	4247	0
JOHN WILEY & SONS	2598	0
ERNST & SOHN-A WILEY CO	1225	0
WILEY-VCH, INC	1142	0
WILEY-V C H VERLAGSGESELLSCHAFT MBH	1100	0
WILEY PERIODICALS	208	0
WILEY EASTERN LTD	36	0
JOSSEY-BASS INC PUBL-JOHN WILEY & SONS	23	0
WILEY CHANCERY	15	0
JOHN WILEY & SONS, LTD,.	0	1
WILEY; COASTAL MORPHOLOGY AND RESEARCH SERIES	0	1
WILEY; SURVEYING AND BOUNDARY CONTROL SERIES	0	1
WILEY/PRAXIS PUBLISHING, CHICHESTER	0	1
ARNOLD, LONDON/WILEY, NEW YORK; UNITED NATIONS ENVIRONMENT PROGRAMME	0	1
WILEY; INTERNATIONAL ASSOCIATION OF GEOMORPHOLOGISTS PUBLICATION, 6	0	1
JOHN WILEY; ECOLOGICAL AND ENVIRONMENTAL TOXICOLOGY SERIES	0	1
JOHN WILEY AND SONS SINGAPORE PTE. LTD.	0	1
WILEY-VCH VERLAG GMBH AND CO. KGAA	0	1
WILEY/PRAXIS PUBLISHING; SERIES IN REMOTE SENSING	0	1
JOHN WILEY AND SONS QUATERNARY RESEARCH ASSOCIATION PROCEEDINGS NO. 5	0	1
WILEY; IUPAC SERIES ON ANALYTICAL AND PHYSICAL CHEMISTRY OF ENVIRONMENTAL SYSTEMS	0	1
RESEARCH STUDIES PRESS LTD, TAUNTON; DISTRIBUTED BY WILEY	0	1
JOHN WILEY; UNESCO ENERGY ENGINEERING SERIES	0	1
WILEY; INTERNATIONAL ASSOCIATION OF GEOMORPHOLOGISTS PUBLICATION, 4	0	1
WILEY; ENVIRONMENTAL SCIENCE AND TECHNOLOGY SERIES	0	1
WILEY; BELHAVEN STUDIES IN CLIMATOLOGY	0	1
EDWARD ARNOLD, LONDON/WILEY, NEW YORK	0	2
WILEY PUBLISHING	0	2
ARNOLD, LONDON/WILEY, NEW YORK	0	4
JOHN WILEY & SONS (ASIA) PTE LTD	0	11
JOHN WILEY AND SONS, INC.	0	13
WILEY-BLACKWELL, LTD	0	25
JOHN WILEY AND SONS, LTD	0	33
WILEY-VCH VERLAG GMBH & CO. KGAA.	0	33
WILEY ONLINE LIBRARY	0	179
WILEY-BLACKWELL PUBLISHING, INC.	0	182
JOHN WILEY AND SONS INC	0	333
JOHN WILEY & SONS, LTD.	0	370
WILEY-ISTE	0	1020
WILEY-BLACKWELL PUBLISHING	0	1361
WILEY-BLACKWELL PUBLISHING ASIA	0	1370
WILEY-VCH VERLAG BERLIN GMBH	0	1592
WILEY-VCH VERLAG GMBH	0	1745
JOHN WILEY AND SONS LTD.	0	2318
WILEY-VCH VERLAG GMBH & CO. KGAA	0	3051
JOHN WILEY & SONS, LTD	0	3959
JOHN WILEY & SONS, INC.	0	4503
WILEY-BLACKWELL PUBLISHING LTD	0	6212
WILEY-VCH	0	6480
WILEY-LISS INC.	0	18120
WILEY BLACKWELL	0	31652
JOHN WILEY AND SONS	0	41402
JOHN WILEY AND SONS LTD	0	68950
JOHN WILEY AND SONS INC.	0	80985
WILEY-VCH VERLAG	0	82931

Abbildung 1: Unterschiedliche Verlagsbezeichnungen für ‘WILEY’ und ihre Auftrittshäufigkeiten im WoS und Scopus

3 Basisdaten

3.1 Entity-Relationship-Modell

Die Basistabellen sollen Informationen zu Verlagen aufnehmen, die für das Zuordnungsverfahren und die spätere Auswertungen notwendig oder hilfreich sind. Diese Informationen werden im Folgenden als Basisdaten bezeichnet. Dazu gehören

- Datumsangaben zu Neugründungen und Schließungen,
- Hierarchiebeziehungen zwischen Einheiten (wie ‚ist Imprint von'),
- Strukturveränderungen (z.B. Verkäufe von Teileinheiten oder Fusionen),
- Namensvarianten (beispielhaft: Bezeichnungen in verschiedenen Sprachen oder Abkürzungen)
- geografische Informationen.[5]

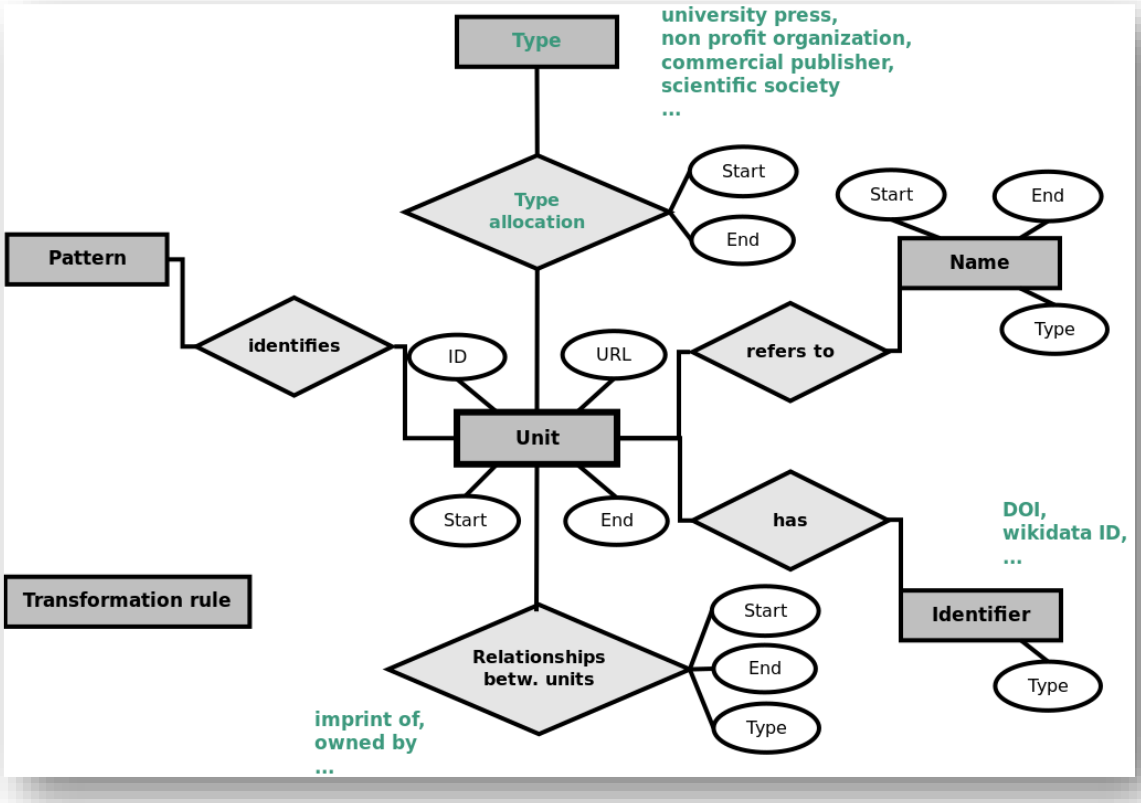


Abbildung 2: ER-Modell für die Basisdaten der verlegerischen Organisationen

[5] Vgl. Chen, Peter P.: *Entity-Relationship Modeling--Historical Events, Future Trends, and Lessons Learned.*

Die Abbildung 2 zeigt das Entity-Relationship-Modell (ER-Modell) mit den Basistabellen. Verlage, Teileinheiten oder übergeordnete Einheiten beliebiger Hierarchieebenen werden als UNIT definiert.

3.2 Initialer Wikidata-Feed

Aufgrund der Vielzahl an Verlagen und dem großen Umfang an Informationen, die zu ihnen und ihren Beziehungen zusammengetragen werden, ist es sinnvoll, diese aus externen Quellen zu beziehen, anstatt sie ausschließlich manuell zu recherchieren. Für die automatische Extraktion von Daten für die Basisinformationen zu den Verlagen wurde als externe Datenquelle Wikidata[6] genutzt.

Wikidata enthält nicht nur die hier interessierenden Daten zu Verlagen, sondern allgemein strukturierte Informationen aus Wikipedia. Diese liegen in Form des Resource Description Framework[7] (RDF) vor und können gut automatisch bearbeitet werden. Im RDF-Format besteht jede Aussage aus einem ‚subject', einer ‚property' und einem ‚value', wobei die vorhandenen Informationen in Form von Tripeln angegeben sind:

< subject > < property > < value >,

wie zum Beispiel:

< Bielefeld_University > < instance_of > < university >.

Für die Information ‚Die Universität Bielefeld ist eine Universität', sind sowohl subject, property als auch value in Form von Identifiern angegeben[8]:

< http : / / www.wikidata.org/entity/Q24382 >

< http : / / www.wikidata.org/prop/direct/P31 >

< http : / / www.wikidata.org/entity/Q3918 >,

wobei die ID Q24382 ‚Bielefeld University', P31 die Property[9] ‚instance_of' und Q3918 ‚University' bezeichnet. Values können sowohl Wikidata-Entitäten (mit zugehöriger ID, wie im angeführten Beispiel) als auch einfache Werte (verschiedener Datentypen wie Strings, Dates, Numbers usw.), wie zum Beispiel das Gründungsdatum sein. Über die Property ‚instance_of' ist eine Klassifikation der Wikidata-Entitäten möglich, wohingegen allerdings nicht für jede Entität ein Triple mit der Property ‚instance_of' vorhanden ist. Zusätzlich

[6] Vgl. Wikidata (2018a)
[7] Vgl. Miller, Eric (1998)
[8] Die Beschreibung der Datenstruktur in Wikidata ist sehr kurz und dient dem groben Überblick. Nicht alle Aussagen sind in dieser einfachen Struktur enthalten, sondern es werden auch sogenannte ‚statements' verwendet. Vgl. MediaWiki (2018).
[9] Für eine vollständige Liste der Wikidata-Properties siehe Wikidata (2018b).

können in bestimmten Fällen für eine Entität mehrere Values für die Property ‚instance_of' vorliegen.

In die Basistabellen des Projekts wurden nur diejenigen Wikidata-Entitäten importiert, die für ‚instance_of' den Value ‚publisher' oder einer Unterkategorie davon aufweisen. Letztere sind gekennzeichnet durch ‚subclass_of'. Offensichtlich nicht benötigte ‚instance_of' Values, wie beispielsweise „Comicverlag", wurden von vornherein ausgeschlossen. Beim Füllen der Basistabellen wurden die Daten bereinigt, wie das folgende Beispiel zeigt:

„2017-01-10T06:51:48Z"^ ^<http://www.w3.org/2001/XMLSchema#dateTime>,

→ 10.01.2017 (im Date-Format der SQL-Datenbank)

oder

„Institut für Bienenkunde Celle"@de
→Institut für Bienenkunde Celle.

3.2.1 Wikidata-Properties und Basistabellen

Um die Informationen aus Wikidata in die Basistabellen aufnehmen zu könnten, ist es nötig, die Wikidata-Properties den entsprechenden Tabellen und Spalten zuzuordnen. Dies geschah mithilfe einer manuellen Recherche und Zuordnung. Abbildung 3 zeigt Beispiele für die vorgenommene Zuordnung von Properties zu Attributen anhand eines Ausschnitts des ER-Modells.

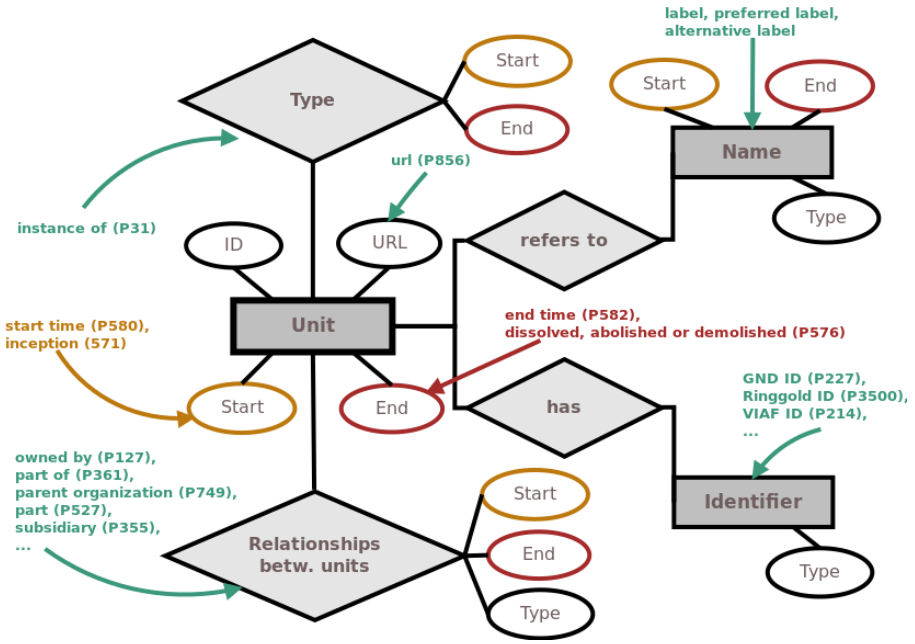


Abbildung 3: Zuordnung der Wikidata-Properties zu den Basisinformationen

In einigen Fällen sind mehrere ausgewählte Properties für ein einziges Attribut relevant. Beispiele dafür sind ‚start_time' und ‚inception' für das Gründungsdatum einer UNIT. Zusätzlich zu den in Abbildung 3 gezeigten Attributen wurden auch probehalber die Rechtsformen und Geo-Informationen (verschiedene Properties wie z. B.: ‚headquarters location', ‚country', etc.) mit aufgenommen.

3.2.2 Coverage initialer Wikidata-Feed

Im Zuge der Übernahme der Informationen in die Basistabellen stellte sich heraus, dass die in Wikidata enthaltenen Verlags-Entitäten nicht vollständig sind. Die im Projekt genutzten Zitationsdatenbanken enthalten weitere Verlagsbezeichnungen, deren zugehörige Entitäten in Wikidata nicht über ‚instance_of' als ‚publisher' klassifiziert sind. Beispiele dafür sind Universitäten, Fachgesellschaften oder Firmen. Diese wurden zusätzlich erfasst. Zudem gibt es auch Fälle, bei denen Verlage nicht in Wikidata enthalten sind. Dies betraf insbesondere Imprints oder andere Teileinheiten. Die entsprechenden Basisdaten konnten daher nicht aus Wikidata automatisch extrahiert werden, sondern wurden manuell aufgenommen.

4 Verfahren: Disambiguierung und Historisierung

4.1 Überblick

In diesem Abschnitt werden die Verfahren dargestellt, mit denen die Verlagsbezeichnungen aus Scopus und dem Web of Science den in den Basistabellen erfassten Verlagsentitäten zugeordnet wurden. Dieser Schritt wird im Folgenden als Disambiguierung bezeichnet. Die Erfassung von Hierarchie- und Strukturveränderungen wird dagegen Historisierung genannt und ebenfalls beschrieben.

Das Vorgehen von Disambiguierung und Historisierung gliederte sich in die folgenden Schritte:

Vorbereitungsschritte: Zur Vorbereitung des Matchings der Verlagsbezeichnungen aus den Bibliometriedatenbanken mit den aus Wikidata stammenden Verlagsentitäten kamen zwei Transformationsvarianten zum Einsatz.

Automatisches Matching: Im Anschluss daran erfolgte ein automatisches Matching mittels String-Matching-Methoden, wobei die Verlagsbezeichnungen aus der WoS- und Scopus-Datenbank den Wikidata-Entitäten in den Basistabellen zugeordnet wurden. Dieser Schritt bedurfte einer sukzessiven Erweiterung, um bisher nicht erfasste Wikidata-Entitäten in die Basistabellen aufzunehmen.

Matching über Textmuster: Für die nicht vom automatischen Matching erfassten Verlagsbezeichnungen, wurden Textmuster angelegt und angewendet.

Manuelle Überprüfung: Bei ausgewählten Einzelfällen erfolgte eine Komplettüberprüfung der Hierarchiebeziehungen und Strukturveränderungen unter Hinzuziehung externer Quellen, wie beispielsweise Verlagswebseiten, deren Archive oder anderen Webseiten.

Anschließend wurden die Basistabellen mit den Wikidata-Entitäten, die durch die Aufnahme von Hierarchiebeziehungen oder Imprints sowie durch die Erstellung von Textmustern neu hinzugekommen sind, befüllt. Dieser Schritt bedurfte einer mehrmaligen Wiederholung, damit für jede manuell aufgenommene Verlagsbezeichnung die zugehörige Relation (parent = Mutterkonzern oder child = Imprint) in den Basistabellen abgebildet wird.

Die Ergebnisse beider Matchingverfahren wurden mithilfe von Prioritätsregeln abschließend zusammengeführt, wobei die Zuordnung auf der niedrigsten Hierarchieebene erfolgte.

Hierarchiebeziehungen: Um die heutige Struktur von Verlagen abzubilden wurde eine Aggregation von Einheiten vorgenommen. Diese Zuordnung bildet dabei die zum Zeitpunkt des Projekts bestehende Relation von Verlagen und Imprints ab, oder, sofern der Verlag nicht mehr existiert, die Struktur zum Zeitpunkt seiner Schließung.[10] Abbildung 4 veranschaulicht die einzelnen Arbeitsschritte.

[10] Die im Projekt abgebildeten Hierarchiebeziehungen entsprechen damit dem Modus A der Institutionenkodierung, die für die Bibliometriedatenbanken des Kompetenzzentrums Bibliometrie bereitgestellt werden. Vgl. Winterhager, M., Schwechheimer, H., & Rimmert, C. (2014).

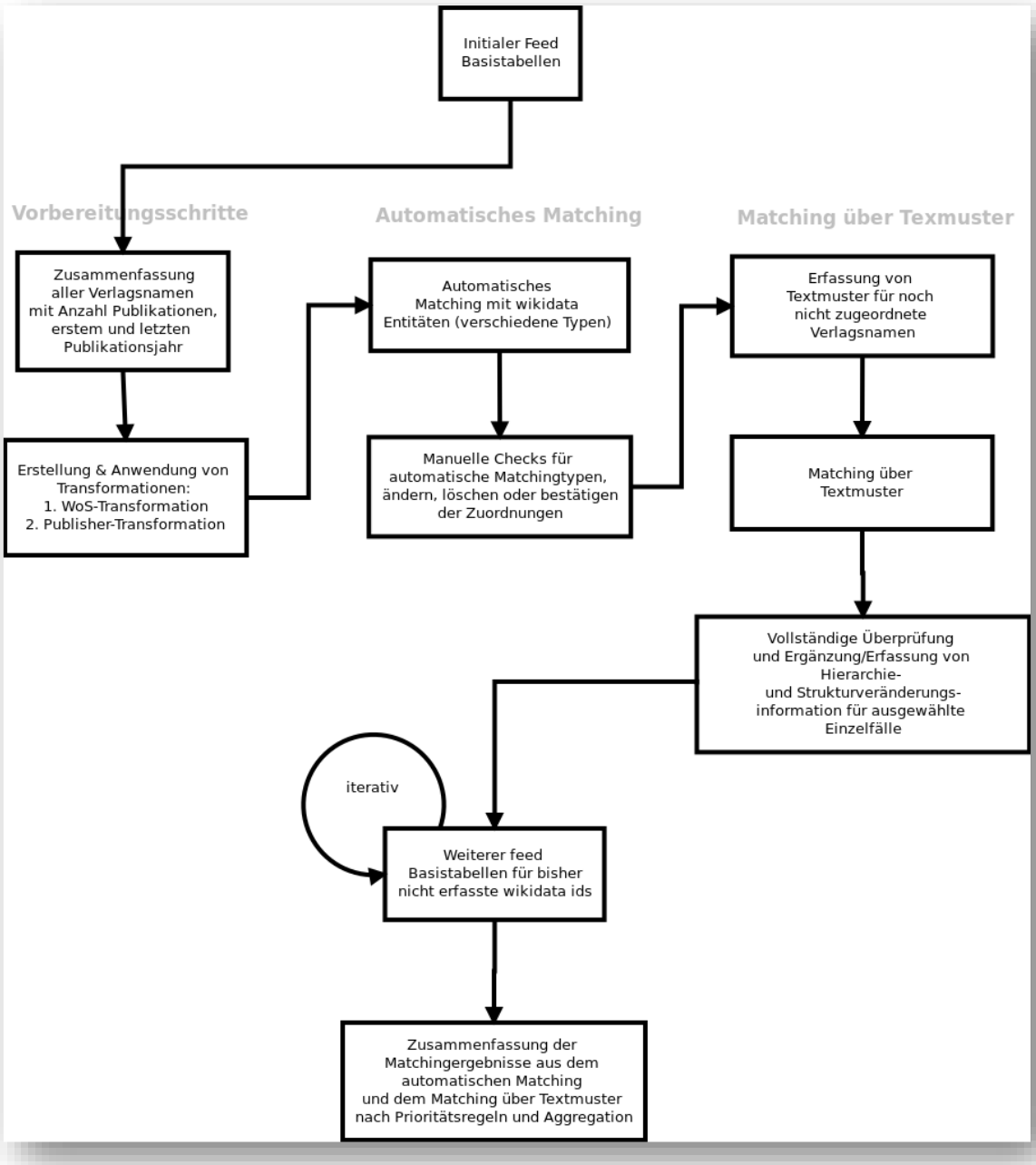


Abbildung 4: Schema des Disambiguierungs- und Historisierungsverfahrens

4.2 Datentransformation

Der Begriff Datentransformation bezeichnet eine Vorstandardisierung sowohl der Verlagsbezeichnungen aus dem Web of Science und Scopus als auch der Namensvarianten der Wikipedia-Entitäten. Mit ihr werden Sonderzeichen gelöscht oder ausgetauscht, sowie Wörter oder eine Kombination aus Wörtern durch einen standardiserten String ersetzt. Der

Ausdruck ‚Universität' („University", „Universidad"), wurde beispielsweise durch ‚UNIV' standardisiert.

Eine Vorstandardisierung ist aus zweierlei Gründen erforderlich: Erstens führt sie zu einer Reduktion der Namensvarianten aus den WoS- und Scopus-Datenbanken. Zweitens sind Verlagsbezeichnungen nur im WoS vorstandardisiert, während dies für die Namensvarianten aus Wikidata und den Verlagsbezeichnungen aus Scopus nicht gilt. Die Transformation auf beiden Seiten erhöht den Recall des automatischen Matchings.

Zur Datentransformation wurden zwei Verfahren angewandt:

1. Die im Rahmen der Institutionenkodierung[11] erstellte Transformation von Adressdaten eignet sich auch für die Transformation von Verlagsbezeichnungen und wurde nachgenutzt.
2. Eine auf den Verlagsfall zugeschnittene Variante. Mit ihr werden die für Verlagsbezeichnungen typische die Stopwords (z. B.: Verlag, Inc., Publishing Company, GmbH usw.) gelöscht und unterschiedliche die Verlagsbezeichnungen unter einem Namen zusammengefasst (Beispiel: SPRINGER und SPRINGER VERLAG → werden unter SPRINGER zusammengefasst.)

4.3 Automatisches Matching

Ziel des automatischen Matching ist es, möglichst viele Verlagsbezeichnungen aus der WoS- und Scopus-Datenbank den Wikidata-Entitäten zuzuordnen. Die Abbildung 5 verdeutlicht, dass es nicht zielführend ist, das Matching auf der kompletten Wikidata-Datenmenge durchzuführen, da es zu vielen unerwünschten Zuordnungen kommen würde. Beispielsweise würde „Springer" nicht nur Verlagen, sondern auch Tieren oder Fahrzeugen zugeordnet werden. Durch eine geeignete Eingrenzung der Zuordnungsziele kann diese Art von Fehler verringert werden.

[11] Vgl. Winterhager, M., Schwechheimer, H., & Rimmert, C. (2014).

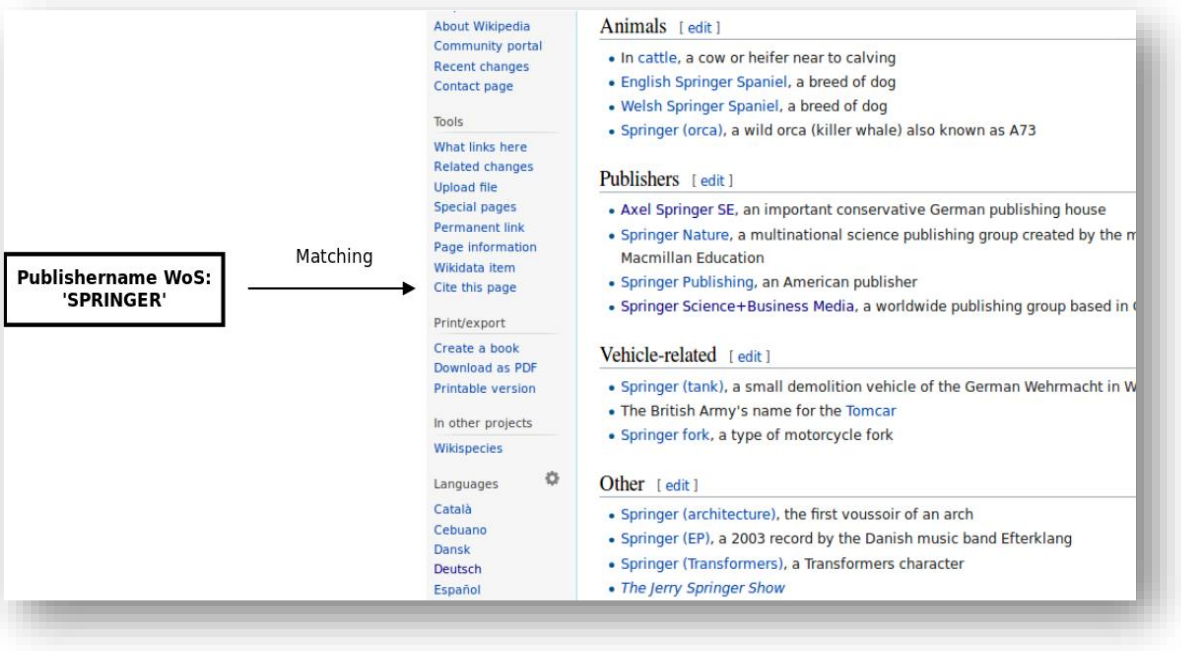


Abbildung 5: Matching auf den gesamten Wikidata-Bestand am Beispiel "Springer“

Eine Eingrenzung der Zuordnungsziele ist allerdings nicht unproblematisch, da die Wikidata-Klassifikation keine Klasse kennt, in der ausschließlich für die Wissenschaft verlegerisch tätige Organisationen zusammengefasst sind. So gibt es zwar mit ‚instance_of' ‚publisher' eine Klasse von Verlagen (welche auch verwendet wurde), diese beinhaltet allerdings auch Comic- oder Spieleverlage, die im vorliegenden Fall nicht interessieren. Daneben finden sich in den WoS- und Scopus-Daten auch Bezeichnungen, die in Wikidata anderen Klassen zugeordnet sind (z. B.: ‚organization', ‚university' und ‚association'). Abbildung 6 stellt einige dieser Fälle dar.

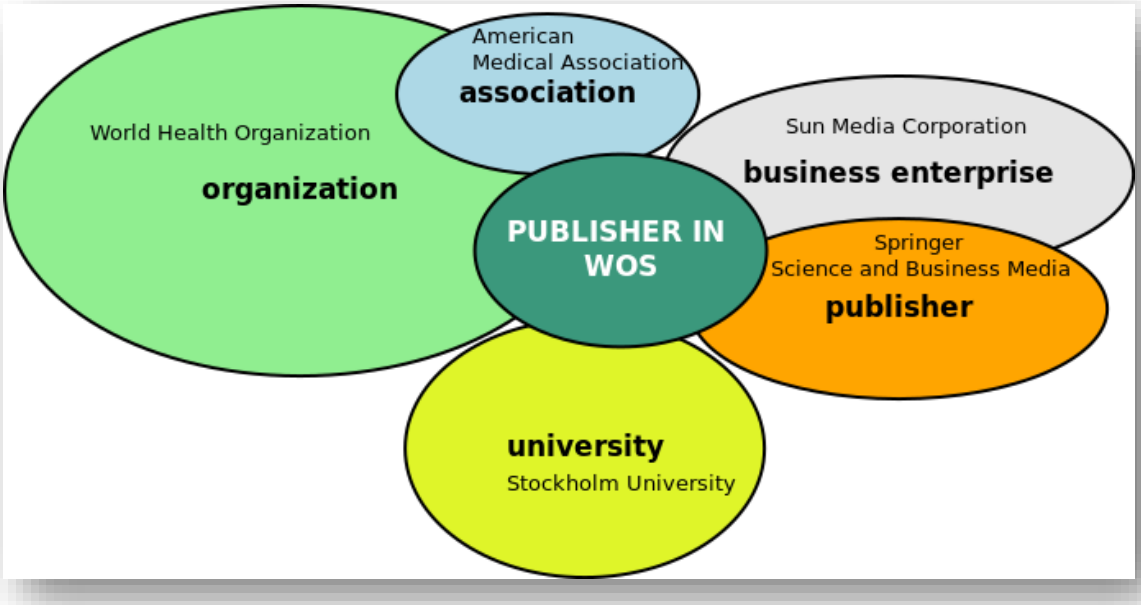


Abbildung 6: ‘Instance_of’-Werte für Verlagsbezeichnungen aus den WoS- und Scopusdaten

Um diesem Problem zu begegnen und einen Überblick darüber zu gewinnen, welche ‚instance_of’-Kategorien zusätzlich zu der Kategorie ‘publisher’ mit berücksichtigt werden müssen, erfolgte zunächst ein Matching auf der gesamten Wikidata-Menge, wobei nur genaue Treffer (die Verlagsbezeichnung aus den WoS- und Scopus-Datenbanken stimmt mit dem Wikidata-Label exakt überein) zugelassen wurden.

Die häufigsten Treffer wurden in vier Kategorien unterteilt, wobei die erste Kategorie ‚instance_of’-Werte enthält die am relevantesten sind und die vierte Kategorie diejenigen zusammenfasst, die eine geringe Relevanz aufweisen (siehe Abbildung 7).

Stufe 1	Stufe 2
RESEARCH INSTITUTE UNIVERSITY ACADEMY OF SCIENCES	BUSINESS ENTERPRISE ASSOCIATION SOCIETY

Stufe 3	Stufe 4
MINISTRY STATISTICAL SERVICE LIBRARY ASSOCIATION	MOUNTAIN CITY CHEMICAL ELEMENT

Abbildung 7: Beispiele verschiedener ‘instance_of’-Werte nach Kategorien

Für die publisher-Entitäten (aus dem initialen Feed) und für die Entitäten mit ‚instance of'-values der Stufen 1 und 2 wurde für das Matching neben der genauen Übereinstimmung von Verlagsbezeichnung mit Wikidata-Namensvariante auch ein Matching über Substrings (Namensvariante ist in Verlagsbezeichnungen enthalten) und Ähnlichkeiten (Namensvariante und Verlagsbezeichnung haben eine Jaro-Winkler-Similarity ≥ 95) durchgeführt. Für die Stufen 3 und 4 wurden hingegen ausschließlich genaue Übereinstimmungen zugelassen und die Zuordnung erfolgte auch erst nach manueller Überprüfung.

Bei dieser Vorgehensweise verblieben Verlagsbezeichnungen zu Wikidata-Entitäten in seltenen/unerwarteten, aber relevanten Kategorien und Wikidata-Entitäten ohne ‚instance of'-value. Um diese durch eine automatische Zuordnung erfassen zu können, wurde das Matching auf Wikidata-Entitäten ausgeweitet, die nicht zu einer als relevant erkannten ‚instance of'-Kategorie gehören, deren Name/Label in Wikidata aber vermuten lässt, dass es sich um relevante Entitäten handeln könnte. Beispiele sind Label/Namensvarianten in denen die Worte Verlag, Publisher, Publishing Company und University vorkommen.

Nach diesem Matching verblieben Verlagsbezeichnungen zu Entitäten, die in Wikidata vorhanden sind, aber bisher nicht einbezogen wurden weil sie keiner als relevant erkannten ‚instance of'-Kategorie angehören und durch ihren Namen/Label nicht als solche erkannt wurden. Auf dieser Restmenge der bislang nicht automatisch zugeordneten Verlagsbezeichnungen wurde ein Matching auf der Wikidata Gesamtmenge durchgeführt und vollständig manuell überprüft, um die Gefahr von Fehlzuordnungen auszuschließen

4.4 Textmuster

Im Fall von Verlagsbezeichnungen, bei denen keine automatische Zuordnung möglich war, fand eine Zuordnung über manuelle Textmuster statt. Hierbei wurden zwei Arten von Textmustern (regulärer Ausdruck oder ein einfaches LIKE-Muster[12]) auf verschiedene Felder (Verlagsbezeichnungen und Adresse, im Fall von WoS zusätzlich auch ‚City') nach verschiedenen Transformationen (4.2) oder auf den Originalwert (ohne Transformation) angewendet. Die Adresse wurde nur hinzugezogen, wenn die Zuordnung auf Grundlage der Verlagsbezeichnung nicht möglich war.

Textmuster können aus zweierlei Gründen erforderlich sein:

- Der Verlag zur Verlagsbezeichnung existiert in Wikidata, konnte aber nicht automatisch zugeordnet werden,
- Der Verlag zur Verlagsbezeichnung existiert in Wikidata nicht.

Im ersten Fall wurde die Wikidata-ID in der Textmustertabelle hinterlegt, da die Basisinformationen über einen weiteren feed automatisch in die Basistabellen mit aufgenommen

[12] http://sqldocu.com/five/likeoperator.htm

wurden. Ein analoges Vorgehen fand bei der Erfassung von Hierarchiebeziehungen und Strukturveränderungen statt, bei denen die zugehörigen child und/oder parent UNITs in Wikidata existieren).

Im zweiten Fall musste der Verlag manuell in die Basistabelle aufgenommen und anschließend mit einem Textmuster versehen werden. Dabei wurde mindestens eine ID vergeben sowie der Name, die Hierarchiebeziehungen und die Strukturveränderungen erfasst.

4.5 Spezielle UNITs

Für spezielle Fälle, bei denen zum Beispiel eine eindeutige Zuordnung des Verlages nicht möglich war oder es sich gar nicht um einen verlegerische Organisation handelte, wurden manuell gesonderte Identifier vergeben. Zu diesen Fällen gehören:

1. Keine Zuordnung möglich: Auf Grund unzureichender oder unvollständiger Verlagsbezeichnungen konnte in einigen Fällen keine eindeutige Zuordnung vorgenommen werden (z.B. weil kein zugehöriger Verlag gefunden werden konnte). Diesen Fällen wurde die V_ID 999999 zugewiesen.

2. Journale: Es kommen nicht nur Verlage in den Verlagsbezeichnungen vor, sondern auch Journale. Hierbei wurde auf eine Verlagszuordnung verzichtet, wobei Informationen des aktuell publizierenden Verlages in einer zusätzlichen Tabelle bereitstehen. Kennzeichnend für Journale ist die V_ID 999998.

3. Collections: In vielen Ländern existieren Fachgesellschaften, die jeweils nur ein Journal herausgeben, in Wikidata nicht vorhanden sind und in einigen Fällen über keine eigene Webseite verfügen. Um den Aufwand für die Einzelaufnahme einer UNIT zu verringern, wurden UNITs angelegt, die die Fachgesellschaften ganzer Länder enthalten. Innerhalb dieser Sammelkategorie wurde je Fachgesellschaft genau ein regulärer Ausdruck als Textmuster angelegt, so dass bei Bedarf eine Einzelzuordnung aufgrund der regulären Ausdrücke vorgenommen werden kann.

4.6 Hierarchien & Strukturveränderungen

In diesem Abschnitt erfolgt die Darstellung der Strukturveränderungen und Hierarchiebeziehungen von Verlagen. Sofern in Wikipedia ein Eintrag für einen Verlag vorlag, konnten alle verfügbaren Informationen zu Hierarchiebeziehungen und Datumsangaben extrahiert und in die jeweiligen Basistabellen überführt werden. Erfolgte eine Zuordnung über Textmuster, wurden Hierarchiebeziehungen und Strukturveränderungen mit erfasst.

Für die fünf größten Verlage – RELX GROUP (ehemals Reed Elsevier), Springer Nature, John Wiley & Sons, Informa Taylor & Francis, Wolters Kluwer – wurden externer Quellen[13] ausgewertet und Informationen zu allen Teileinheiten und Strukturveränderungen aufgenommen.

Strukturveränderungen sind bei größeren Verlagen häufig zu beobachten und oftmals sehr komplex. Eine Schwierigkeit besteht darin, geeignete Informationsquellen zu diesen Veränderungen zu finden, insbesondere dann, wenn sie länger zurückliegen. Die beiden Abbildungen 8 und 9 illustrieren dies am Beispiel der Strukturveränderungen und Hierarchiebeziehungen von J.B. Lippincott & Co[14].

History [edit]

The publisher had its origins in a Philadelphia bookstall opened by Benjamin Warner and Jacob Johnson in 1792. Joshua Ballinger Lippincott assumed control of the firm in 1836. In 1978, the company (then named J. B. Lippincott Company) was sold to Harper & Row, at which point it began to focus its publishing activities exclusively in health care; in 1990, it was sold to Wolters Kluwer. It was later merged with Raven Press in 1995 to become **Lippincott-Raven Publishers**, which then merged with Williams & Wilkins, ultimately forming **Lippincott Williams & Wilkins** in 1998.

Abbildung 8: Strukturveränderung in Wikipedia, Beispiel J.B. Lippincott & Co

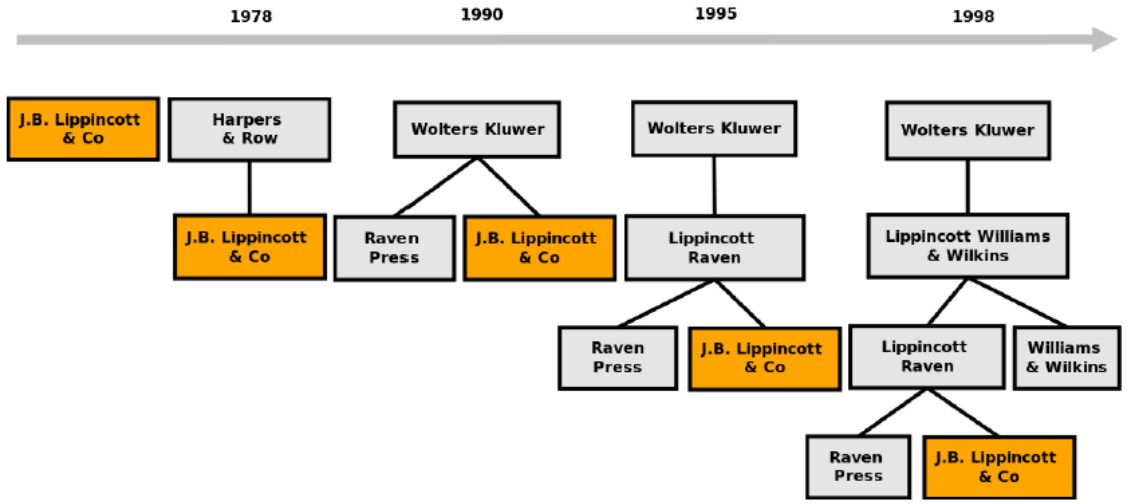


Abbildung 9: Strukturveränderungen, Beispiel J.B. Lippincott & Co

[13] Wie z. B. das News Archiv von Taylor & Francis, Taylor & Francis Group (2018)
[14] Vgl. Wikipedia (2018a)

4.7 Zusammenfassung der Matching-Ergebnisse und Aggregation

Die aus dem automatischen und dem textmusterbasierten Matching-Verfahren gewonnenen Zuordnungen wurden nach folgenden Prioritätsregeln übernommen:

1. Manuell geprüfte Zuordnungen (aus den Matchings auf publisher-Entitäten und über die Stufen, Jaro-Winkler-Similarity ≥ 95 oder wikidata-Namensvariante als Substring in der Verlagsbezeichnung enthalten)
2. Zuordnungen über Textmuster
3. Zuordnungen aus den Ergebnissen der Überprüfung von Mehrfachzuordnungen (Beschreibung folgt in 5.3.8)
4. Zuordnungen zu Publisher-Entitäten, genaue Treffer
5. Zuordnungen zu Entitäten mit ‚instance of'- values der Stufe 1
6. Zuordnungen zu Entitäten mit ‚instance of'- values der Stufe 2
7. Zuordnungen zu Entitäten relevantem Label (z.B. 'UNIV')
8. Manuell geprüfte Zuordnungen (aus den Matchings auf Wikidata gesamt)
9. Zuordnungen mit einer Jaro-Winkler-Similarity von ≥ 95 und Wikidata-Namensvariante ist als Substring in der Verlagsbezeichnung enthalten.

Bei den ersten Tests wurden zunächst alle Zuordnungen beibehalten, um über entstehende Mehrfachzuordnungen Hinweise auf Fehlerquellen zu erhalten. Bei der abschließenden Zuordnung wurde das Zuordnungsverfahren Schritt für Schritt durchlaufen und jeder Schritt jeweils nur noch auf die Restmenge der noch nicht zugeordneten Verlagsbezeichnungen angewandt. Für die Zuordnungen von Verlagsbezeichnungen zu Identifiern von Verlags-Entitäten wurde anschließend nur die Zuordnung zu den UNITs der niedrigsten Hierarchieebene beibehalten. Anschließend wurden die Verlagsentitäten aggregiert. Es wurden dabei die heute anzutreffende oder im Fall von nicht mehr existierenden Verlagen die zuletzt existierende Hierarchiebeziehungen abgebildet. Für die Bestimmung der niedrigsten Hierarchieebene als auch die Aggregation wurden dabei nur Beziehungen vom Typ 2 verwendet (vergleiche dazu 5.1.1).

4.8 Check von Mehrfachzuordnungen

Nach einem ersten Durchlauf des Zuordnungsverfahrens wurden die dabei erzielten Mehrfachzuordnungen, also Verlagsnamen, die mehreren Verlagen zugeordnet sind, manuell überprüft. Diese wurde auf Verlagsbezeichnungen beschränkt, die im WoS auftreten, denen in den letzten 10 Jahren mindestens 100 Publikationen/Dokumenten zugeordnet sind.

Mehrfachzuordnungen können verschiedene Ursachen haben:

- In einer Verlagsbezeichnung sind verschiedene Verlage enthalten, die Mehrfachzuordnung ist korrekt.
- Es liegt ein Zuordnungsfehler vor, beispielsweise bedingt durch Verlage mit gleichen/-ähnlichen Namensvarianten in Wikidata. Hier wurden Zuordnungen aus dem automatischen Zuordnungsverfahren ausgeschlossen oder Textmuster korrigiert.
- Es wurden für einen Verlag mehrere Identifier vergeben (Dubletten). Bei diesem Fehler wurde nur eine Zuordnung beibehalten.
- Es fehlen Angaben zu Hierarchiebeziehungen/Strukturveränderungen. Die entsprechenden Angaben wurden ergänzt.

Da die Basisdaten aus Wikidata bezogen wurden, führen Dubletten in Wikidata zu Dubletten in den Basistabellen (für den Fall, dass beide in die Basistabellen aufgenommen wurden)

5 Ergebnisse

Eine vollständige Zuordnung sämtlicher Verlagsbezeichnungen aus den beiden Bibliometriedatenbanken war im Rahmen dieses Kleinprojekts nicht möglich. Daher wurde folgender Schwerpunkt gesetzt: Zugeordnet wurden Verlagsbezeichnungen, für die im Publikationszeitraum 2008-2017 mindestens 100 Publikationen nachgewiesen wurden.

5.1 Tabellenbeschreibung

Ausgeliefert wurden die aus Wikidata extrahierten und manuell ergänzten Basisinformationen zu Verlagen in Form von Basistabellen, die auf Grundlage des ER-Modells (Abb.2) erstellt wurden sowie Tabellen mit Zuordnungen der Verlagsbezeichnungen zu den für die Verlage vergebenen Identifiern. Die Tabellen werden im Folgenden näher beschrieben. Abbildung 10 zeigt das zugehörige Tabellenschema.

5.1.1 Basistabellen

Zentral ist die Tabelle ‘V UNIT’, in der für jeden Verlag ein Identifier verzeichnet ist (Spalte ‘V ID’). Die Wikidata-ID eignet sich nicht als Identifier, da zum einen nicht für

alle Verlage eine Wikidata-ID vergeben wurde, während für andere Verlage gleich mehrere Wikidata-IDs existieren. Zudem enthält die Tabelle die Spalte URL sowie FIRST- und LASTDATE um Gründungs- und Schließungsdaten aufzunehmen.

In ‘V NAME’ sind für jeden Verlag eine oder mehrere Namensvariante/n verschiedener Typen erfasst. Auch für Namen sind FIRST- und LASTDATE vorhanden, um Namenswechsel erfassen zu können. Um sich je Verlag (UNIT) und Zeitpunkt nur einen Namen anzeigen zu lassen, kann der Typ 5 verwendet werden (bei Namenswechseln sind mehrere Namensvarianten mit Typ 5 vorhanden). Die Lookup-Tabelle für Namenstypen ist die Tabelle ‘V NAME TYPE’. Die hier verwendeten Typen gehen z.T. auf Wikidata zurück (wie ‚label’ oder ‚preferred label’).

Die Tabelle ‘V IDENTIFIER’ enthält verschiedene Typen von Identifiern (Lookup-Tabelle für Identfier-Typen ist ‘V IDENTIFIER TYPE’). Diese sind in den Spalten ‘V IDENTIFIER’ ‘TYPE ID’ und ‘ID VALUE’ erfasst. Die Tabelle ‘V CLASSIFICATION’ ist für Klassifikationen von Verlagen vorgesehen und kann mehrere Klassifikationssysteme aufnehmen. Lookup-Tabelle für Klassifikationstypen (‘V CLASSIFICATION TYPE ID’) ist ‘V CLASSIFICATION TYPE’.

Die Tabelle ‘V RELATION’ enthält Beziehungen verschiedener Typen zwischen UNITs. Dabei sind bisher 3 Typen vergeben worden: Typ 1 (‘owned by’-property aus Wikidata), Typ 2 (eine Zusammenfassung mehrere properties aus Wikidata, die Hierarchiebeziehungen darstellen) und Typ 3 (weiterer Typ ohne Bezug zu Wikidata-Properties – dieser Typ ist als ‘other’ bezeichnet, in der bereitgestellten Version enthält er aber nur Beziehungen zwischen einem Verlag und einer Holding). Eine Holding soll üblicherweise nicht als ‘parent’/Hauptverlag des Verlags erscheinen, die Beziehung kann aber, je nach Kontext, von Interesse sein.

Um Strukturveränderungen abzubilden sind in der Tabelle ‘V RELATION’ die Spalten FIRST- und LASTDATE vorhanden.

Die Tabelle ‘V GEO UNITS’ enthält die in Wikidata vorhandenen geografischen Informationen zu Verlagen, die testweise extrahiert wurden. Die verschiedenen Typen (wie zum Beispiel ‘country’, ‘headquarters location’ oder ‘postal code’) sind in der Spalte ‘GEO TYPE’ zu finden.

Entitäten, die in WoS oder Scopus nicht vorkommen oder solche, die über Relationen mit Verlagsbezeichnungen aus den Bibliometriedatenbanken verbunden sind, wurden in den Basistabellen nicht gelöscht, um möglichst viele Informationen zu erhalten. Um aus der Tabelle ‘V UNIT’ nur diejenigen Verlage zu erhalten, die im WoS und/oder Scopus tatsächlich vorkommen, muss über die jeweilige Zuordnungstabelle entsprechend eingeschränkt werden.

5.1.2 Zuordnungstabellen

Die Zuordnungen der Verlagsbezeichnungen aus den Bibliometriedatenbanken zu real existierenden Verlagen

Verlagsbezeichnung -> V ID

werden in zwei Varianten bereitgestellt: Zum einen wird die Zuordnung auf der niedrigst möglichen Hierarchieebene angegeben. Diese Zuordnung ermöglicht (in Kombination mit der Tabelle ‘V RELATION’) eine individuelle Aggregation je nach Projektanforderung – inklusive Einzelfallentscheidungen für bestimmte Verlage.

Zusätzlich wird (als eine mögliche Form der Aggregation) der Hauptverlag nach den Regeln des Modus A angegeben. Es werden also die entweder anzutreffenden gültigen oder – im Fall von nicht mehr bestehender UNITs – die zuletzt existierenden Hierarchiebeziehungen verwendet.

Zur Zuordnung auf ‘Hauptverlagsebene’ wird außerdem in der Spalte ‘PATH’ der Pfad (beginnend bei der zugeordneten V ID niedrigster Hierarchieebene und endend beim ‘Hauptverlag’) für die Aggregation angegeben, über den die hierarchische Struktur für die betreffende UNIT nachvollzogen werden kann.

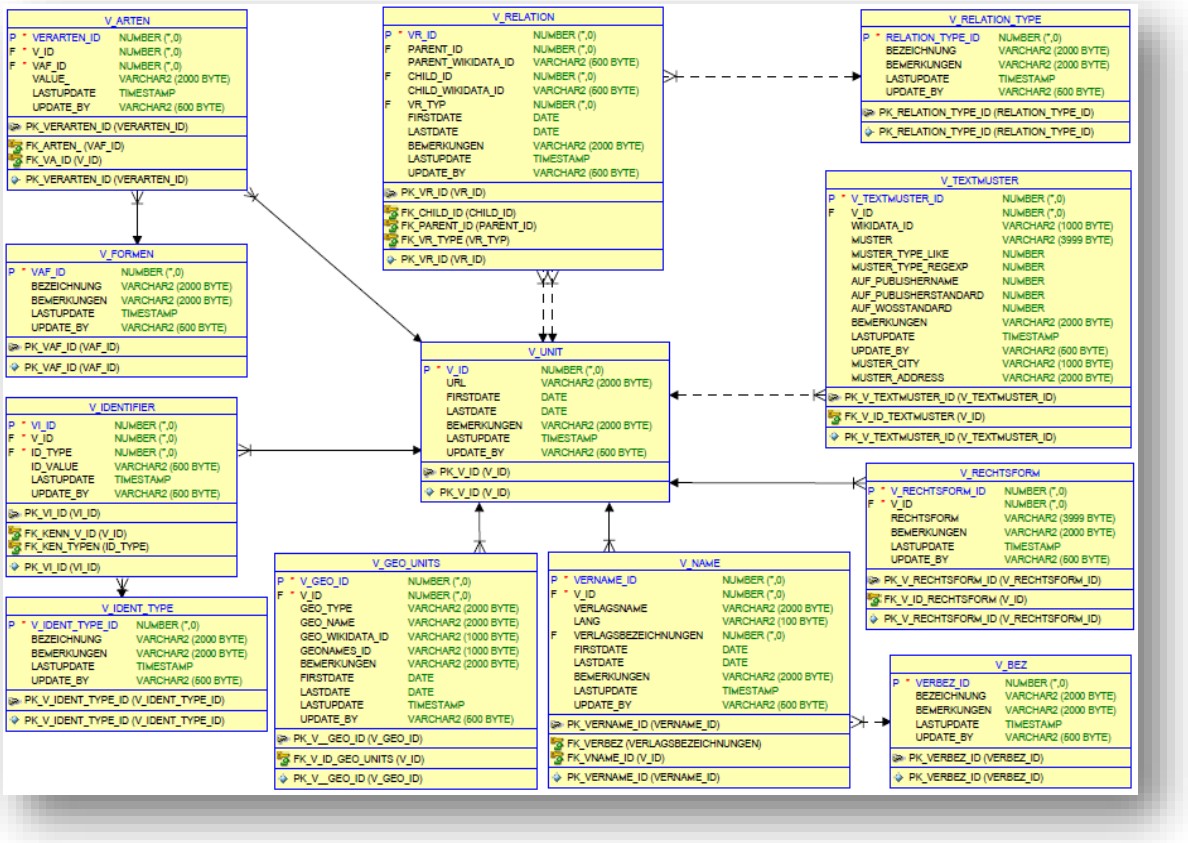


Abbildung 10: Tabellenschema der Basistabellen des Projektes "Publisher: Disambiguierung und Historisierung"

5.2 Statistik

5.2.1 Basisinformationen

Die folgende Tabelle zeigt die Anzahl der aufgenommenen UNITs und Beziehungen, aufgeschlüsselt nach der Anzahl der automatisch übernommenen und manuell ergänzten Informationen:

	Gesamt	Wikidata-Feed	manuelle Aufnahme
Anzahl UNITs	11.548	10.358	1.190
Anzahl relations	1.177	845	332

Tabelle 2: UNITs/relations Abdeckung des Wikidata-Feed

Von den aufgenommenen Beziehungen zwischen den Entitäten sind dabei 168 vom Typ 1 (‘owned by’ property), 950 vom Typ 2 (Teileinheit) und 39 vom Typ 3 (Holding).

Der Umfang der für Verlagsentitäten vorhandenen Informationen geht aus der folgenden Tabelle 3 hervor.

Attribut	#UNITs, für die das Attribut vorhanden ist
Geo-Iinformationen	7.958
Identifier	10.466
Identifier ohne Wikidata-ID	3.825
Typ (‚instance of’ property)	9.975

Tabelle 3: UNITs für vorhandene Attribute

Geo-Iinformationen fasst verschiedenen Arten von Daten zusammen, darunter headquarters location, postal code, country u.ä.

Die Top 10 ‘instance of’-Werte sind:

1. PUBLISHER (5516)
2. UNIVERSITY (1180)
3. ORGANIZATION (573)
4. BUSINESS ENTERPRISE (472)
5. SCIENTIFIC JOURNAL (195)
6. UNIVERSITY PRESS (195)
7. PUBLIC EDUCATIONAL INSTITUTION OF THE UNITED STATES (182)
8. RESEARCH INSTITUTE (168)
9. GOVERNMENT AGENCY (115)
10. PRIVATE NOT-FOR-PROFIT EDUCATIONAL INSTITUTION (114)

In Klammern ist dabei die Anzahl der UNITs angegeben, die der entsprechenden Klasse angehören. Zu beachten ist dabei, dass eine UNIT mehreren ‚instance of’-Werten zugeordnet sein kann und dass durch die iterativen Wikidata-Feeds parents und childs von Verlagsentitäten aufgenommen wurden, die nicht unbedingt wieder Verlagsentitäten sein müssen (im Fall von SCIENTIFIC JOURNAL beispielsweise bestehen in einigen Fällen

‚owned by' Zuordnungen von Verlagen zu Journalen). PUBLISHER ist hier zwar die häufigste Klasse, aber es ist deutlich erkennbar, dass auch andere Klassen relevant sind.

Für das Disambiguierungsverfahren wurden insgesamt 6.024 Textmuster erstellt (LIKE-Muster und reguläre Ausdrücke).

5.2.2 Disambiguierung

Die folgende Tabelle zeigt die Zuordnungsquote von distinkten Verlagsbezeichnungen für den Gesamtzeitraum (WoS: ab 1980, Scopus: ab 1995) je Datenbank und für Verlagsbezeichnungen, die in der entsprechenden Datenbank innerhalb der letzten 10 Jahre vorkommen[15].

Mit 61% für WoS und 46% für Scopus ist der Anteil der zugeordneten Verlagsbezeichnungen für den Gesamtzeitraum relativ niedrig (Tab. 4). In der höheren Zuordnungsquote von 91% für die vergangenen 10 Jahre im Fall des WoS zeigt sich der im Rahmen dieses Projekts gelegte Schwerpunkt (Tab. 5).

	#Verlagsbezeichnungen	davon zugeordnet	in %
WoS	11.815	7.174	60,72
Scopus	13.047	5.972	45,78

Tabelle 4: Zugeordnete Verlagsbezeichnungen in WoS und Scopus ab 1980

	#Verlagsbezeichnungen mit PY in WoS/Scp ≥ 2008	davon zugeordnet	in %
WoS	5.077	4.637	91,33
Scopus	8.762	4.569	52,14

Tabelle 5: Zugeordnete Verlagsbezeichnungen in WoS und Scopus mit Pubyear ≥ 2008

Die Betrachtung des Anteils der zugeordneten Verlagsbezeichnungen ist jedoch nur bedingt aussagekräftig, da häufig und selten Verlagsbezeichnungen als gleichwertig behandelt werden. Da im Projektkontext und auch im Fall einer Nachnutzung in der Regeln nicht

[15] Die Anzahl der Verlagsbezeichnungen wurde hier jeweils ohne Berücksichtigung der Adresse bestimmt. Es existieren Verlagsbezeichnungen, für die in Kombination mit mehr als einer Adresse vorkommen (mehrere PK PUBLISHERS je PUBLISHERNAME).

die distinkten Verlagsbezeichnungen, sondern die einem Verlag zugeordneten Publikationen von primärem Interesse sind, ist es sinnvoll, die Zuordnungsquote anhand von letzterem zu bestimmen. Dazu zeigt die folgenden Tabellen die Anzahl der Dokument-Verlagsbezeichnungs-Kombinationen und den jeweils zugeordneten Anteil, für den Gesamtzeitraum sowie für die letzten 10 Jahre:

	#Dokument-Verlagsbez.-Kombinationen	davon zugeordnet	in %
WoS	53.496.843	50.003.920	93,47
Scopus	11.643.905	10.406.458	89,73

Tabelle 6: Dokument-Verlagsbez.-Kombinationen für WoS und Scopus

	#Dokument-Verlagsbez.-Kombinationen mit PY in WoS/Scp ≥ 2008	davon zugeordnet	in %
WoS	42.070.852	41.890.302	99,57
Scopus	11.180.232	10.095.655	90,30

Tabelle 7: Dokument-Verlagsbez.-Kombinationen mit PY in WoS/Scp ≥ 2008 für WoS und Scopus

Der Anteil der zugeordneten Publikationen ist mit 93% für WoS und 89% für Scopus deutlich höher als der Anteil der zugeordneten distinkten Verlagsnamen. Auch hier zeigen sich in den höheren Anteilen für Publikationsjahre ab 2008 sowie für das WoS der hier gesetzte Schwerpunkt.

Bei der Zuordnung der Verlagsbezeichnungen des WoS erfolgten 10% über automatisch generierte und manuell überprüfte Treffer (Scopus: 2%) und 58% über Textmuster (Scopus 54%) und der verbleibende Anteil über verschiedene automatische Matchings mit den in 4.8 angegebenen Prioritäten.

Für das WoS wurde aus den Zuordnungen eine Zufallsstichprobe von 100 Verlagsbezeichnungen (aus dem Gesamtzeitraum) mit zugeordnetem Verlag und Hauptverlag manuell geprüft. Eine Zuordnung wurde dabei als korrekt eingestuft, wenn sowohl die Zuordnung auf der niedrigsten Hierarchieebene als auch die Zuordnung zum Hauptverlag korrekt ist. Die folgende Tabelle zeigt das Ergebnis der Überprüfung:

	Anzahl	Fälle
Korrekte Zuordnung	94	
Fehler	4	1. drei Fälle von Verlagen/Institutionen mit gleichem Namen in anderem Land und ein Fall einer Mehrfachzuordnung - diese Zuordnung ist falsch, die richtige Zuordnung ist aber auch vorhanden
Spezialfälle	2	2. ein Fall unklar und in einem Fall Teileinheitenzuordnung falsch, aber parent richtig (anderer Campus der gleichen Universität)

5.3 Herausforderungen und Besonderheiten

In diesem Teil werden einige Herausforderungen des Projekts vorgestellt und Lösungsmöglichkeiten aufgezeigt.

5.3.1 Longtail

In beiden Datenbanken gibt es eine große Anzahl seltener Verlagsbezeichnungen. Für einen besseren Überblick über die Verlagsbezeichnungen mit Auftrittshäufigkeiten ≤ 100 zeigen die folgenden Tabellen die Verteilung.

Auftrittshäufigkeit WoS	#Verlagsbezeichnungen	in %
≤ 100	3.622	30,66
≤ 10	10.577	4,88
= 1	102	0,86

Tabelle 8: Auftrittshäufigkeiten der Verlagsbezeichnungen im WoS

Auftrittshäufigkeit Scp	#Verlagsbezeichnungen	in %
≤ 100	7.794	59,74
≤ 10	3.347	25,65
= 1	1.774	13,60

Tabelle 9: Auftrittshäufigkeiten der Verlagsbezeichnungen im Scopus

Im WoS treten insgesamt 31% der Verlagsbezeichnungen über den Gesamtzeitraum maximal hundertmal auf, während die Auftrittshäufigkeit solcher seltenen Verlagsbezeichnungen in Scopus bei 60% liegt.

5.3.2 Identifier für Verlage

Die Suche nach Identifiern für Verlage, die das Matching erleichtern könnten, blieb erfolglos. In den Bibliometriedatenbanken sind Digital Object Identifier (DOI) für Publikationen vorhanden, wobei ein Teil der DOI den Verlag kennzeichnet (DOI-Präfix[16]).

Ein Ansatz für eine Zuordnung bestand darin, die mit dem Dokument verlinkte Verlagsbezeichnung über das Dokument mit dem DOI-Präfix zu verknüpfen und Verlagsbezeichnungen zusammenzufassen. Dies scheiterte zum einen daran, dass für viele Verlage kein Präfix existiert, zum anderen traten viele offensichtliche Fehler auf (Zusammenfassungen von Bezeichnungen verschiedener Verlagsentitäten, unerwünschte Mehrfachzuordnungen). Ein Problem sind dabei möglicherweise Strukturveränderungen. Eine Verlagsbezeichnung, die ein Imprint von Verlag 1 identifiziert, wird über ein oder mehrere Dokumente einem bestimmten Präfix zugewiesen. Die gleiche Verlagsbezeichnung kann, sofern der Verlag zum Imprint eines anderen Verlags wird, über andere Dokumente mit einem ganz anderen Präfix verlinkt sein.

5.3.3 Kooperation von Gesellschaften und großen Verlagen

Größere Verlage kooperieren häufig mit Fachgesellschaften, so dass sich hinter einer Verlagsbezeichnung ‘Society XY’ auch beispielsweise ein Verlag wie Wiley verbergen kann[17]. Zu den auf diese Weise publizierten Journalen finden sich auf den entsprechenden Webseiten Hinweise wie ‚published by ... on behalf of …’. In diesem Fall könnte, je nach Fragestellung, sowohl eine Zuordnung zum Verlag als auch zu der Fachgesellschaft sinnvoll sein. Eine genaue Abgrenzung zwischen der verlegerisch tätigen Organisation und dem Inhaber des Zeitschriftentitels konnte im Rahmen des Projekts nicht vorgenommen werden. Dies würde den Rechercheaufwand immens erhöhen und die Prüfung, erfordern, ob und wenn ja, wie sich die Kooperation von Fachgesellschaft und Verlag im Laufe der Zeit verändert. Grundsätzlich können solche Kooperationen im Datenschema aber erfasst werden, etwa durch die Definition eines neuen Relationstypen in der Tabelle V RELATION.

5.3.4 Relationstypen

Typ 1 erwies sich als schwer zu interpretieren. Die ‘owned by’-property wurde mehrmals für Inhaber der Verlage oder auch Holdings vergeben, manchmal aber auch im Sinne einer Teileinheit verwendet. Aufgrund dieser Interpretationsschwierigkeiten wurde der Relationstyp 1 zwar mit aufgenommen, jedoch nicht in der Aggregation verwendet. Vor Verwendung dieser Information ist daher eine manuelle Überprüfung und Bereinigung empfehlenswert.

[16] Vgl. Wikipedia (2018b)
[17] Vgl. Wiley

Für Holdinggesellschaften (die sowohl unter den parents mit Typ 2 als auch unter parents mit Typ 1 zu finden sind), wurde ein weiterer Relationstyp (3) vergeben. Einige Einzelfälle wurden manuell in diesen Typ geändert bzw. manuell zusätzlich aufgenommen. Außerdem wurden alle Typen von Relationen, bei denen der parent entweder im Label den Substring 'HOLDING' enthält oder über 'instance of' der Klasse 'holding company' (Q219577) angehört, in Typ 3 geändert. Es ist nicht klar, ob damit alle Holdinggesellschaften als Typ 3 erfasst wurden. Sofern unter Typ 2 weitere Holdinggesellschaften vorhanden sind, können diese nach Aggregation unerwünscht als 'Hauptverlag' auftauchen. Solche Fälle ließen sich über den in der Zuordnungstabelle angegebenen Aggregationspfad bereinigen.

Durch die Aufnahme der drei hier genannten Relationstypen sind nicht alle möglichen Relationen zwischen Verlagsentitäten abgedeckt. Die Relationstabellen sind darauf ausgelegt, bei Bedarf weitere Relationstypen und Relationen aufzunehmen. Diese könnten beispielsweise aus relevanten Wikidata properties stammen, aus anderen externen Quellen erschlossen oder manuell erfasst werden.

5.3.5 Journale

In den Verlagsbezeichnungen sind zum Teil auch Journalnamen vertreten. Diesen Fällen wurde der Identifier V ID 999998 zugewiesen. In einigen Fällen konnte der Verlag recherchiert werden, der zum Recherchezeitpunkt das Journal publiziert. Da sich die Journal-Verlag-Beziehung ändern kann, wurden auch die Verlagsbezeichnungen der V ID 999998 und nicht dem recherchierten Verlag zugeordnet. Die Information zum Verlag wurden aber in einer zusätzlichen Tabelle ('V JOURNAL INFOS') zur Nutzung mit erfasst.

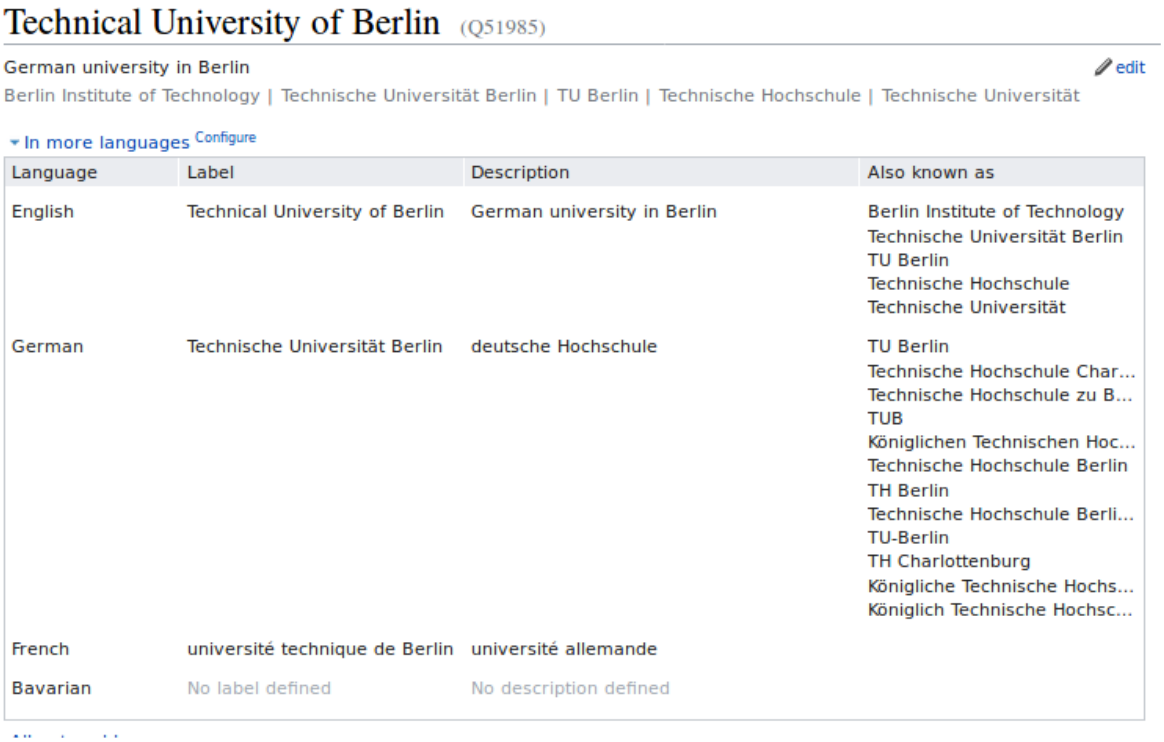
5.3.6 Relationen

Für die Aggregation auf Hauptverlagsebene werden Hierarchiebeziehungen benötigt. Diese wurden über bestimmte Wikidata properties identifiziert.

Dabei entstehen zwei Probleme: Zum einen werden die Wikidata properties nicht immer mit der gleichen Intention verwendet. So sind unter den zur property ‚owned 'by' angelegten Beziehungen sowohl Hierarchiebeziehungen als auch Beziehungen anderer Art wie beispielsweise Firmeninhaber zur Firma zu finden. Zudem sind nicht immer alle relevanten Hierarchiebeziehungen in Wikidata erfasst, so dass manuelle Ergänzungen erforderlich sein können. Hier ist u.a. auch die Definition der Hauptverlagsebene entscheidend für die Frage, welche Entitäten aufgenommen und verknüpft werden müssen. Dies kann sich je nach Projektkontext unterscheiden. '

5.3.7 Nicht eindeutige Namensvarianten in Wikidata

Einige Fälle von Wikidata-Namensvarianten sind nicht eindeutig. Abbildung 11 zeigt das Beispiel der TU Berlin in Wikidata, für die eine Namensvariante ‘Technische Universität’ vorhanden ist, die nach automatischer Transformation ,‘TECH UNIV’) - zu vielen unerwünschten Treffern führt. Diese fielen in der Überprüfung der Mehrfachzuordnungen (siehe 4.9) auf und konnten bereinigt werden.



Technical University of Berlin (Q51985)

German university in Berlin

edit

Berlin Institute of Technology | Technische Universität Berlin | TU Berlin | Technische Hochschule | Technische Universität

In more languages Configure

Language	Label	Description	Also known as
English	Technical University of Berlin	German university in Berlin	Berlin Institute of Technology Technische Universität Berlin TU Berlin Technische Hochschule Technische Universität
German	Technische Universität Berlin	deutsche Hochschule	TU Berlin Technische Hochschule Char... Technische Hochschule zu B... TUB Königlichen Technischen Hoc... Technische Hochschule Berlin TH Berlin Technische Hochschule Berli... TU-Berlin TH Charlottenburg Königliche Technische Hochs... Königlich Technische Hochsc...
French	université technique de Berlin	université allemande	
Bavarian	No label defined	No description defined	

Abbildung 11: Wikidata Namensvarianten zur Technischen Universität Berlin

Literaturverzeichnis

Chen, Peter P.: *Entity-Relationship Modeling--Historical Events, Future Trends, and Lessons Learned.* [Online verfügbar unter:] https://bit.csc.lsu.edu/~chen/pdf/Chen_Pioneers.pdf (letzter Zugriff am: 03.05.2018)

Kompetenzzentrum Bibliometrie (2018): Über das Kompetenzzentrum Bibliometrie. [Online verfügbar unter:] http://www.bibliometrie.info/ (letzter Zugriff am 03.05.2018)

MediaWiki (2018): File statements. [Online verfügbar unter:] https://www.mediawiki.org/wiki/Help:File_statements (letzter Zugriff am 03.05.2018)

Miller, Eric (1998): An Introduction to the Resource Description Framework. In: D-Lib Magazine, May 1998, ISSN: 1082-9873, [Online verfügbar unter:] http://www.dlib.org/dlib/may98/miller/05miller.html (letzter Zugriff am: 03.05.2018)

Taylor & Francis Group (2018): LIBRARIANRESOURCES. Taylor & Francis supporting librarians. Library Insights Blog. [Online verfügbar unter:] http://www.tandf.co.uk/libsite/news/newsArchive/news2003.asp (letzter Zugriff am: 03.05.2018)

Wikipedia (2018a): Artikel über Lippincott Williams & Wilkins [Online verfügbar unter:] https://en.wikipedia.org/wiki/Lippincott_Williams_%26_Wilkins (letzter Zugriff am 03.05.2018)

Wikipedia (2018b): Digital Object Identifier. [Online verfügbar unter:] https://de.wikipedia.org/wiki/Digital_Object_Identifier (letzter Zugriff am 03.05.2018)

Wikidata (2018a): Main Page. [Online verfügbar unter:] https://www.wikidata.org/wiki/Wikidata:Main_Page (letzter Zugriff am 03.05.2018)

Wikidata (2018b): Wikidata, List of Properties, [Online verfügbar unter:] https://www.wikidata.org/wiki/Wikidata:List_of_properties (letzter Zugriff am: 03.05.2018)

Wiley (2018): Societies. [Online verfügbar unter:] https://www.wiley.com/en-us/societies (letzter Zugriff am 03.05.2018)

Winterhager, M., Schwechheimer, H., & Rimmert, C. (2014): Institutionenkodierung als Grundlage für bibliometrische Indikatoren. *Bibliometrie - Praxis und Forschung*, *3*(14), S. 1 - 22.