Universität Bielefeld
Institute for Interdisciplinary Studies of Science (I²SoS)

kompetenzzentrum bibliometrie

Machbarkeitsstudie

Institutionen-Kodierung Dimensions

Abschlussbericht

16.10.2020

Universität Bielefeld
AG Bibliometrie
PD Dr. Niels Taubert
Christopher Lenke
Postfach 10 01 31
33501 Bielefeld

E-mail:
niels.taubert@uni-bielefeld.de
christopher.lenke@uni-bielefeld.de

1. Einleitung

In den bislang für bibliometrische Analysen am häufigsten genutzten Zitationsdatenbanken Web of Science (WoS) und Scopus finden sich neben anderen Informationen auch die institutionellen Adressen von den an Publikationen beteiligten AutorInnen. Die Qualität dieser Informationen lässt hinsichtlich der Genauigkeit, Vollständigkeit und Standardisierung zu wünschen übrig. Um hier Abhilfe zu schaffen und um eine eindeutige Zuordnung von Adressen zu real existierenden Forschungseinrichtungen in Deutschland vorzunehmen, hat die Arbeitsgruppe Bibliometrie ein entsprechendes Verfahren für die beiden Datenbanken entwickelt. Bis einschließlich zum Jahr 2020 wurde die AG Bibliometrie damit beauftragt, den Partnern des Kompetenzzentrums Bibliometrie (KB) jeweils für die prozessierten Rohdatenbanken und die Bibliometrie-Datenbanken des WoS und Scopus Tabellen bereitzustellen, die eine Adresszuordnung erlauben. Die Institutionen-Kodierung stellt dabei zwei Zuordnungsvarianten bereit: Eine Zuordnung der Adressen zu der aktuellen (im Fall von geschlossenen Einrichtungen: zur zuletzt existierenden) Institution (Modus A) oder zu der Institution, wie sie sich zum Publikationszeitpunkt darstellte (Modus S) (Winterhager et al. 2014; Rimmert et al. 2017).

In der jüngeren Vergangenheit sind mehrere Datenquellen entstanden, die auch für bibliometrische Analysen interessant sind. Verglichen mit den beiden genannten Datenbanken liegen bislang weit weniger Erfahrungen vor und Umfang und Qualität der darin enthaltenen Daten sind weit weniger gut bekannt. Eine dieser Datenquellen ist die Datenbank Dimensions,[1] deren Lizensierung seitens der Konsortialpartner in Vorbereitung auf den Nachfolgeantrag KB 2021+ diskutiert wird. Sie verfügt ebenso wie das WoS und Scopus über eine Adresszuordnung. Zu Vollständigkeit und Qualität dieser Informationen können bislang keine belastbaren Aussagen getroffen werden.

Vor dem Hintergrund der skizzierten Entwicklungen und des Diskussionsstands innerhalb des KB ist es Ziel des Kleinprojekts, die Machbarkeit einer Anwendung der Institutionen-Kodierung auf die Datenbank Dimensions zu prüfen und die Ergebnisse der Anwendung zu evaluieren. Es zielt darauf ab, Hintergrundinformationen für die Entscheidung zu liefern, ob eine Lizensierung der Datenbank Dimensions im Rahmen des Fortsetzungsantrags KB 2021+ sinnvoll ist und ob eine Institutionen-Kodierung für diese Datenbank angeboten werden sollte. Im Einzelnen soll untersucht werden,

(a) zu welchen Zuordnungsergebnissen die Anwendung der Institutionen-Kodierung auf die Datenbank Dimensions führt,
(b) ob die Anwendung der Institutionen-Kodierung auf die Datenbank Dimensions zu einem Mehrwert gegenüber dem nativen Institutionen-Identifier der Datenbank (GRID_ID[2]) führt,
(c) mit welchen Aufwänden finanzieller und personeller Art eine regelmäßige Erstellung der Institutionen-Kodierung für die Datenbank Dimensions verbunden ist.

2. Methode

Das Vorgehen der Machbarkeitsstudie setzt sich aus den folgenden drei Schritten zusammen:

[1] https://www.dimensions.ai/.

[2] GRID_ID ist der Institutionen-Identifier der Global Research Identifier Database (GRID). Neben dem Identifier stellt die Datenbank weitere Informationen zu den darin enthaltenen Institutionen, wie Adresse, Webseite und Geo-Informationen bereit.

2.1 Vorbereitende Evaluierung

Die Institutionen-Kodierung nutzt eine Reihe von Informationen zur Zuordnung von Adressvarianten zu einer real existierenden Institution. Im Fall von Dimensions sind das die Felder:

- ADDRESS_FULL: Adress-String aus der Tabelle ITEMS_AUTHORS_INSTITUTIONS[3]
- VENDOR_ITEM_ID: Publikations-Identifier aus ITEMS
- GRID_ID: Institutionen-Identifier aus ITEMS_AUTHORS_GRID
- COUNTRY_CODE: Länderkennung der Adresse gemäß ISO 3166-1 alpha 2 aus ADDRESSES
- CITY: Name der Stadt aus ADDRESSES[4]
- PUBYEAR: Publikationsjahr aus ITEMS.

In einem ersten Schritt wurde die Vollständigkeit der Informationen in der Datenbank Dimensions geprüft. Dieser Schritt dient als Grundlage, um die weiteren Ergebnisse der Machbarkeitsstudie einordnen zu können.

2.2 Analyse des Adress-Strings ADDRESS_FULL

Die Institutionen-Kodierung ist bislang auf das Web of Science und Scopus hin optimiert und berücksichtigt die jeweiligen Spezifika der Adress-Strings der Datenbanken. Es ist damit zu rechnen, dass die Struktur des Adress-Strings in ADDRESS_FULL von Dimensions mehr oder weniger stark von den beiden anderen Datenbanken abweicht. Daher wurde in einem zweiten Schritt die Struktur des Adress-Strings der Dimensions-Datenbank analysiert und geprüft, ob sie den Anforderungen an einen Input der Institutionen-Kodierung entspricht.

2.3 Evaluation der Zuordnungsergebnisse

Im Anschluss an die beiden vorbereitenden Schritte wurde die Institutionen-Kodierung auf zwei Samples[5] angewendet:

- *Sample Schnittmenge WoS*: Ursprünglich war geplant, ein Sample von 50.000 randomisiert ausgewählten DOI von Publikationen zu bilden, die sowohl im WoS als auch in Dimensions verzeichnet sind, um danach die Institutionen-Kodierung auf sämtliche der in beiden Datenbanken hinterlegten Adress-Dokument-Kombinationen anzuwenden. Danach hätten die für beide Datenbanken erzielten Zuordungsergebnisse verglichen werden sollen. Das Vorgehen scheiterte aber an der Datenstruktur von Dimensions: Im Fall von Items, die eigentlich mit mehreren Adressen hätten verknüpft sein müssen, bildet Dimensions meist nur eine Relation ab. Daher wurde ein alternatives Vorgehen gewählt: Erstellt wurde eine Zufallsstichprobe mit 100.000 deutschen Adress-Dokument-Kombinationen aus der Datenbank Dimensions.[6] Die Adress-Dokument-Kombinationen stammt dabei aus Journalen, die sowohl im WoS als auch in

[3] Bezeichnungen der Tabellen und der Tabellenfelder sind der Dokumentation des Dimensions-Datenbankschemas in Stahlschmidt und Sohn (o.Jg.) entnommen.
[4] ADDRESSES ist Teil des SQL-Schemas von GRID.
[5] Da die AG Bibliometrie selbst über keinen Zugang zur Dimensions Datenbank verfügt, wurden die beiden Samples für die Machbarkeitsstudie am 27.Juni vom DZWH bereitgestellt. Wir danken herzlich für diese Unterstützung.
[6] Da im Rahmen dieses Vorgehens die Auswahlgesamtheit Adress-Dokument-Kombinationen und nicht Publikationen sind, finden sich in dieser Stichprobe nicht sämtliche Adress-Kombinationen aus der Datenbank Dimensions, die für die zugehörigen Publikationen verzeichnet sind.

Dimensions erfasst sind. Aus diesem Sample wurde per Zufallsauswahl eine Stichprobe mit 50.000 distinkten DOI mit sämtlichen im Sample zugehörigen Adress-Dokument-Kombinationen (51.008) gebildet. Diese Stichprobe bildete die Datengrundlage für die Anwendung der Institutionen-Kodierung im Fall von Dimensions. Für die Anwendung der Institutionen-Kodierung auf einen Testdatensatz des WoS wurden sämtliche Adress-Dokument-Kombinationen, die mit denselben 50.000 DOI im WoS verknüpft sind aus der WoS_B_2020 ausgewählt. Da hier sämtliche Adress-Dokument-Kombinationen berücksichtigt werden konnten, war der Testdatensatz im Fall des WoS mit 151.419 distinkte Adress-Dokument-Kombinationen wesentlich größer als der Testdatensatz im Fall von Dimensions.

- *Sample Differenzmenge Non-WoS*: Hier war ebenfalls geplant, ein Sample mit sämtlichen Adressen aus der Datenbank Dimensions zu bilden, die mit 10.000 randomisiert ausgewählten DOI verknüpft sind. Die DOI sollten dabei Publikationen in Journalen identifizieren, die nicht in der Web of Science Master Journal List verzeichnet sind. Auch dieses Vorgehen war aufgrund der oben dargestellten Datenstruktur von Dimensions nicht möglich. Stattdessen wurden 10.000 Adress-Dokument-Kombinationen aus Dimensions randomisiert ausgewählt, die mit Publikationen aus nicht im WoS indexierten Journalen verknüpft sind. Da dieselbe Adresse mehrfach mit derselben Publikation verknüpft sein kann (z.B. in Fällen einer Publikation mit mehreren Autoren aus derselben Einrichtung) oder eine Adresse mit mehreren unterschiedlichen Publikationen verknüpft sein kann, liegt die Anzahl distinkter Adress-Dokument-Kombination bei 9.842 und die Anzahl distinkter Adressen bei 9.182.

Die dabei erzielten Zuordnungsergebnisse wurden in einem dreistufigen Prozess evaluiert:

- *Zuordnungsstatistik:* Aus der Zuordnungsstatistik gehen Kennzahlen wie die prozentuale Zuordnung distinkter Adressen aus der Menge sämtlicher Adressen, sowie die Zuordnungsquote der Adressen in verschiedenen Zeiträumen hervor.
- *Vergleich der Zuordnungsergebnisse der Datenbanken Dimensions und WoS:* Die Zuordnungsergebnisse der Institutionen-Kodierung der Datenbanken Dimensions und WoS wurden auf der Grundlage des ersten Samples miteinander verglichen.[7]
- *Analyse nicht zugeordneter Adressen:* In einem letzten Schritt wurde anhand von 100 nicht zugeordneten Adressen analysiert, aus welchen Gründen keine Zuordnung erfolgte.

3. Ergebnisse

3.1 Vorbereitenden Evaluierung

In einem ersten vorbereitenden Schritt wurde die Vollständigkeit der für die Institutionen-Kodierung notwendigen Information für sämtliche Adressinformationen aus Deutschland geprüft. Diese Prüfung erfolgte auf der gesamten Dimensions-Datenbank und wurde aufgrund des Fehlens eines Datenzugangs bei der AG Bibliometrie vom DZHW mit dem folgenden Ergebnis durchgeführt. Von den 117.367.880 Einträgen verfügen sämtliche über einen Identifier (VENDOR_ITEM_ID). Bei 104.237.447 oder 88,81% dieser Einträge handelte es sich im so genannte ‚aktive Einträge', also solche, die nicht nachträglich entfernt oder

[7] Ursprünglich war geplant, den Anteil der identisch zugeordneten Adressen und Adress-Dokument-Kombinationen für das Sample ‚Schnittmenge WoS' zu bestimmen. Dies war ebenfalls aufgrund der Datenstruktur von Dimensions nicht möglich.

zusammengeführt wurden. Die in Tabelle 1 genannten Kennzahlen beziehen sich auf die Vollständigkeit der Informationen von aktiven Einträgen.

Tabelle 1: Für die Institutionen-Kodierung benötigte Felder: Mit Items verknüpfte Informationen

Feld	*Beschreibung*	*Anzahl*	*Vollständigkeit (%)*
VENDOR_ITEM_ID	Publikations-Identifier der Datenbank Dimensions	104.237.447	100
VENDOR_ITEM_ID mit Autorenangaben	VENDOR_ITEM_ID mit mindestens einer oder mehreren Angabe(n) zu Autoren.	94.573.833	90,73
PUBYEAR	Publikationsjahr	104.215.021	99,97

Neben den genannten Feldern auf der Ebene von Publikationen sind für die Anwendung der Institutionen-Kodierung natürlich auch Adress-Informationen notwendig, die in der Datenstruktur von Dimensions mit Autoren verknüpft sind. In der nachstehenden Tabelle 2 bezieht sich die Vollständigkeit der Felder ADDRESS_FULL und GRID_ID auf die Angaben zu Autoren.

Tabelle 2: Für die Institutionen-Kodierung benötigte Felder: Mit Autoren verknüpfte Informationen

Feld	*Beschreibung*	*Anzahl*	*Vollständigkeit (%)*
Autorenangaben	Anzahl Zeilen in ITEMS_AUTHORS_INSTITUTIONS	310.021.192	100
Autorenangabe mit ADDRESS_FULL	Vorliegen einer ADDRESS_FULL in ITEMS_AUTHORS_INSTITUTIONS (Adress-String) für eine Autorenangabe	201.262.181	64,92
Autorenangaben mit GRID_ID	Vorliegen einer GRID_ID für eine Autorenangabe	155.004.564	50,00

Die Prüfung der Vollständigkeit der benötigen Datenfelder führt zu einem gemischten Ergebnis: Eine gute Abdeckung findet sich für Datenfelder aus der Tabelle ITEMS.

Weniger positiv erscheint dagegen die Vollständigkeit der mit Autoren verknüpften Informationen. Sowohl die Abdeckung der GRID_ID als auch die Vollständigkeit des Feldes ADDRESS_FULL sind als kritisch einzuschätzen: Für ein Drittel der Autoren liegen keine Adressinformationen vor. Um die Bedeutung dieses Desiderats abschätzen zu können, ist an diesem Punkt eine tiefergehende Analyse auf der Ebene der gesamten Datenbank notwendig, mit der u.a. geklärt werden müsste, ob sich die Vollständigkeit von ADDRESS_FULL in Richtung rezenter Publikationsjahrgänge verbessert. Da Dimensions die beiden Datenbankfelder COUNTRY_CODE und CITY aus GRID bezieht, entspricht deren Vollständigkeit der der GRID_ID (50,00%).

3.2 Analyse des Adress-Strings ADDRESS_FULL

Die Analyse der Struktur des Adress-Strings aus ADDRESS_FULL ergab eine prinzipielle Brauchbarkeit als Input für die Institutionen-Kodierung. Eine weitergehende Transformation für eine erstmalige Anwendung der Kodierung erwies sich als nicht notwendig.

3.3 Evaluation der Zuordnungsergebnisse

3.3.1 Sample Schnittmenge WoS

Für das erste Sample ‚Schnittmengemenge WoS' wurden für 50.000 DOI insgesamt 46.027 distinkte zugehörige Adressen in der Datenbank Dimensions ermittelt. Von diesen konnten 43.865 distinkte Adresse zugeordnet werden, was einer Zuordnungsquote von 95,30% entspricht. In ähnlicher Größenordnung bewegen sich die Kennzahlen für die distinkten Adress-Dokument-Kombinationen: Von den 51.008 distinkten Adress-Dokument-Kombinationen konnten 48.704 zugeordnet werden, woraus sich eine Zuordnungsquote von 95,48% ergibt. Die untenstehende Tabelle 3 zeigt dabei an, dass die Zuordnungsquote über die verschiedenen Zeiträume konstant ist.

Tabelle 3: Sample Schnittmenge WoS, Zuordnungsquote der aus Dimensions stammenden Adress-Dokument-Kombinationen

Zeitraum	*Anzahl dist. Adress-Dokument-Kombinat.*	*davon zugeordnet*	*Zuordnungsquote (%)*
2015-2019	17.431	16.615	95,32
2010-2014	17.917	17.115	95,52
2005-2009	11.851	11.321	95,53
2000-2004	3.809	3.653	95,90
Gesamt	51.008	48.704	95,48

Für dieselben 50.000 DOI wurden auf der Grundlage des Web of Science sämtliche zugehörigen Adressinformationen ermittelt und durch die Institutionen-Kodierung zugeordnet. Von den insgesamt 95.585 distinkten Adressen konnten 90.948 zugeordnet werden, was einem Anteil von 95,15% entspricht. Die Anzahl der distinkten Adress-Dokument-Kombinationen ist mit 151.419 noch höher, von denen 146.521 oder ein Anteil von 96,77% zugeordnet werden konnten. Die Tabelle 3 zeigt an, dass die Zuordnungsquote über die hier berücksichtigten Zeiträume ebenfalls konstant ist.

Tabelle 4: Sample Schnittmenge WoS, Zuordnungsquote der aus dem Web of Science stammenden Adress-Dokument-Kombinationen

Zeitraum	*Anzahl dist. Adress-Dokument-Kombinat.*	*davon zugeordnet*	*Zuordnungsquote (%)*
2015-2019	54.458	52.059	95,59
2010-2014	55.492	53.958	97,24
2005-2009	31.836	31.081	97,63
2000-2004	9.633	9.423	97,82
Gesamt	151.419	146.521	96,77

Aus dem Vergleich der Zuordnungsergebnisse der Institutionen-Kodierung für das Sample Schnittmenge WoS können zwei Ergebnisse festgehalten werden.

- Erstens ist die Zuordnungsquote für die Datenbank Dimensions nur geringfügig kleiner als die für das WoS. Dies spricht für eine grundsätzliche Eignung der Institutionen-Kodierung für die Zuordnung von Adressen aus Dimensions zu real existierenden Forschungseinrichtungen.

- Zweitens lässt es die Datenstruktur von Dimensions nicht zu, die Vollständigkeit der Adress-Informationen in ADDRESS_FULL in Dimensions mit ADDRESS_FULL im WoS zu vergleichen, noch den Anteil zugeordneter und nicht zugeordneter Adressen beiden Datenbanken miteinander ins Verhältnis zu setzen. Dies würde eine identische Konstruktion des Samples für beide Datenbanken voraussetzen.

3.3.2 Sample Differenzmenge WoS

Mit dem zweiten Sample ‚Differenzmenge WoS' soll geprüft werden, ob bei Adress-Informationen von Publikationen aus Dimensions, die nicht im Web of Science verzeichnet sind, Zuordnungsquoten in ähnlicher Größenordnung erzielt werden. Für dieses zweite Sample ‚Differenzmenge WoS' wurden für 10.000 DOI insgesamt 9.182 distinkte zugehörige Adressen in der Datenbank Dimensions ermittelt. Von diesen konnten insgesamt 8.726 distinkte Adressen zugeordnet werden, was einer Zuordnungsquote von 95,03% entspricht. In ähnlicher Größenordnung bewegen sich die Kennzahlen für die distinkten Adress-Dokument-Kombinationen: Von den 9.842 distinkten Adress-Dokument-Kombinationen konnten 9.365 zugeordnet werden, woraus sich eine Zuordnungsquote von 95,15% ergibt. Die untenstehende Tabelle 5 zeigt dabei an, dass die Zuordnungsquote über die verschiedenen Zeiträume konstant ist.

Tabelle 5 Sample Differenzmenge WoS, Zuordnungsquote der aus Dimensions stammenden Adress-Dokument-Kombinationen

Zeitraum	*Anzahl dist. Adress-Dokument-Kombinat.*	*davon zugeordnet*	*Zuordnungsquote (%)*
2015-2019	5.996	5.702	95,10
2010-2014	2.913	2.785	95,61
2005-2009	740	707	95,54
2000-2004	192	170	88,54
Gesamt	9.842	9.365	95,15

Die Zuordnungsquote von Adress-Informationen zu Publikationen, die zwar in Dimensions nicht aber im WoS verzeichnet sind, bewegt sich in einer ähnlichen Größenordnung wie die Zuordnungsquote von Adress-Information zu Publikationen, die im WoS verzeichnet sind.

Auch diese Ergebnisse erbringen Evidenz für die grundsätzliche Eignung der Anwendung der Institutionen-Kodierung auf Adress-Informationen aus Dimensions. Ein relativer Ausreißer ist dabei die Zuordnungsquote für Adress-Dokument-Kombinationen aus den Jahren 2000-2004. Zu berücksichtigen sind allerdings hier die vergleichsweise geringe Fallzahl, die dazu führt, dass bereits eine einzelne zugeordnete bzw. nicht zugeordnete Publikation zu einer deutlichen Veränderungen der Zuordnungsquote führt.

3.3 Analyse nicht zugeordneter Adressen:

Um das Potential für eine Qualitätsverbesserung der Institutionen-Kodierung für die Datenbank Dimensions abzuschätzen wurde für ein Sample von jeweils 100 Fällen intellektuell geprüft, weswegen eine Zuordnung nicht erfolgte.

Die Ursachen für eine nicht erfolgende Zuordnung von Adress-Informationen zu real existierenden Institutionen können vielfältig sein. Tabelle 5 gibt eine Übersicht über die Ursachen sowie Beispiele und die Häufigkeit des Auftretens. In der letzten Spalte wird eine Einschätzung vorgenommen, ob das Problem durch eine automatisierte Datenbereinigung vor Anwendung der Institutionen-Kodierung beseitigt werden kann. Bei der Ursache ‚Fehlen eines Textmusters' ist zu beachten, dass die Ursache Gegenstand der permanenten Qualitätsverbesserung der Institutionen-Kodierung ist und durch die Anlage neuer Institutionen und neuer Textmuster, sowie der Anpassung bestehender Textmuster kontinuierlich bearbeitet wird.

Tabelle 6: Ursachen, Häufigkeit und Lösbarkeit einer fehlenden Zuordnung von Adress-Informationen in Dimensions zu real existierenden Institutionen

Ursache	*Beispiel*	*Häufigkeit*	*Lösbar**
Fehlen eines Textmusters		78	(Ja)
Sonderzeichen, UTF-8 Kodierung (Umlaute)	Univ Düsseldorf an Stelle von Univ Dusseldorf	14	Ja
Sonderzeichen ISO 8859-15 Kodierung (u.a. französische)	É	8	
Unzulässige Interpunktion/Sonder-zeichen im Adressfeld	Semikolon, Fragezeichen, Klammer, Anführungszeichen	--	Ja
Bei den Informationen handelt es sich um keine Adresse	E-Mailadressen, Telefonnummern und Homepages im Adressfeld, Autoreninformation	--	Nein
Deutsche Adresse für eine ausländische Institution	Biotechnology Institute Thurgau (BITg) at the University of Konstanz, Unterseestrasse 47, CH-8280 Kreuzlingen, Switzerland University of Konstanz, Universitätsstrasse 10, Konstanz 78464, Germany	--	Nein
Wechselnde Sprachen im Adressfeld	From: Abteilung Kardiologie und Angiologie […] and Abteilung Nephrologie (F.H.B., H.H.), Medizinische Hochschule Hannover, Hannover, Germany.	--[8]	(Nein)

* Gemeint ist hier die Lösbarkeit durch automatisierte Datenbereinigung.

Aus der Tabelle 6 geht hervor, dass die drei am häufigsten auftretenden Ursachen für eine fehlende Zuordnung grundsätzlich lösbar sind. Die häufigste Ursache ist das Fehlen eines Textmusters, mit der die Zuordnung einer Adressvariante zugeordnet werden kann, gefolgt von zwei Arten von Problemen mit Sonderzeichnen. Letztgenannte Ursachen können durch eine Anpassung der Datenbereinigungs-Prozeduren zur Vorbereitung der Institutionen-Kodierung beseitigt werden.

4. Schlussfolgerungen

Das Projekt ‚Machbarkeitsstudie Institutionen-Kodierung Dimensions' konnte nicht sämtliche in der Projektskizze formulierte Ziele erreichen. Die Gründe dafür liegen in spezifischen Eigenschaften der Datenbank, die als Defizite charakterisiert werden müssen:

[8] Die vier in der Tabelle zuletzt genannten Ursachen einer fehlenden Zuordnung traten nicht innerhalb des Samples auf, sondern fielen im Verlauf des Projekts bei der Analyse des Adress-Strings auf (siehe 3.2).

- Erstens bildet die Datenbank im Fall von Items, die eigentlich mit mehreren Adressen verknüpft sein müssten, meist nur eine Relation ab. Daher konnte die Stichprobenziehung nicht wie geplant erfolgen und musste durch ein alternatives Verfahren realisiert werden. Dies schloss allerdings einen Teil der im Rahmen des Projekts anvisierte Analyseschritte aus. Dazu zählt der Vergleich von identisch und nicht identisch zugeordneten Adressen bei der Anwendung der Institutionen-Kodierung auf Dimensions gegenüber dem Web of Science. Über das Projekt hinaus ist daher festzuhalten, dass die Nutzbarkeit der Datenbank Dimensions in der vorliegenden Form durch die mangelnde Abbildung von Item-Adress-Relationen erheblich eingeschränkt ist.
- Zweitens ist die Vollständigkeit der Felder ADDRESS_FULL (64,92%) sowie GRID_ID (50,00%) in der gesamten Datenbank als problematisch zu bewerten. Mangelnde Vollständigkeit kann dabei zwei Ursachen haben. Im Fall von ADDRESS_FULL kann entweder die Information in dem Tabellenfeld schlicht fehlen, oder die Adressverknüpfung ist in der Tabelle ITEMS_AUTHORS_INSTITUTIONS nicht angelegt (komplette Zeile fehlt). Analoges gilt für die GRID_ID und die mit ihr verbundenen Felder COUNTRY_CODE und CITY. Welche der beiden Ursachen dafür verantwortlich ist, ist auf der Grundlage der beiden uns vorliegenden Sample nicht zu klären.
- Drittens setzt die Anwendung der Institutionen-Kodierung das Vorhandensein sämtlicher unter 2.2 genannten Tabellenfelder voraus, darunter die Felder CITY und COUNTRY_CODE. Da diese aber aus der Tabelle ADDRESSES stammen und damit zwingend auch mit einer GRID_ID verknüpft sind, konnte die Zuordnungsleistung der Institutionen-Kodierung nur einseitig bestimmt werden. Praktisch ermittelt wurde die Zuordnungsleistung der Institutionen-Kodierung gegenüber Adressen, die in Dimensions zugeordnet sind. Nicht bestimmt werden konnte hingegen die Zuordnungsleistung der Institutionen-Kodierung gegenüber Adressen, die in Dimensions nicht zugeordnet sind. Daher kann auch die Frage nach einem möglichen Mehrwert der Institutionen-Kodierung für Adressen der Datenbank Dimensions gegenüber der Zuordnung durch den Institutionen-Identifier GRID_ID nicht beantwortet werden.

5. Empfehlungen

Aus den im Rahmen der Machbarkeitsstudie gewonnenen Erfahrungen mit der Datenbank Dimensions resultieren die folgenden Empfehlungen:

1. Eine Lizensierung der Datenbank Dimensions kann vor dem Hintergrund der im Rahmen der Studie erzielten Befunde nur empfohlen werden, sofern die Item-Adress-Relationen deutlich vollständiger abgebildet werden als es bislang der Fall ist und das Feld ADDRESS_FULL zumindest in aktuellen Jahrgängen weitgehend vollständig ist.
2. Weniger schwer wiegt die geringe Vollständigkeit von CITY und COUNTRY_CODE. Bei weitgehender Vollständigkeit der Item-Adress-Relationen sowie der Informationen in ADDRESS_FULL (siehe oben) können die beiden Felder aus dem Adress-String extrahiert und für die Durchführung der Institutionen-Kodierung verwendet werden. Ein solches Vorgehen wäre der möglichen Alternative einer Durchführung der Institutionen-Kodierung ohne die Nutzung der beiden Felder (angesichts zu erwartender Laufzeitprobleme und mangelnder Interpretierbarkeit der dabei entstehenden Zuordnungsstatistik) vorzuziehen.

3. Bei Vorliegen der notwendigen Felder führt die Anwendung der Institutionen-Kodierung auf Dimensions zu brauchbaren Zuordnungsergebnissen. Die Zuordnungsquote entspricht in der Größenordnung der des WoS.
4. Die Ursachenanalyse für eine fehlende Zuordnung zeigt, dass das Potential der Institutionen-Kodierung noch nicht vollständig ausgeschöpft ist. Durch eine automatische Bereinigung des Adress-Strings kann eine Erhöhung der Zuordnungsquote um ~1% realisiert werden. Nach einer Anpassung der Institutionen-Kodierung (Entwicklung eines Verfahrens zur Extraktion des Adress-Strings und automatische Bereinigungsschritte, ca. 1 PM) würde der Aufwand für Erstellung und Qualitätsverbesserung der Institutionen-Kodierung den Aufwänden für die beiden anderen Datenbanken entsprechen.

Literatur

Rimmert C, Schwechheimer H, Winterhager M. *Disambiguation of author addresses in bibliometric data-bases - technical report.* Bielefeld: Universität Bielefeld, Institute for Interdisciplinary Studies of Science (I²SoS); 2017.

Stahlschmidt S, Sohn A. Dimensions Raw Data. SQL DB Service. Schema Documentation: dim20190926, Version 20200425. DZWH.

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