

Bericht Funding Acknowledgement-Datenbereinigung Web of Science

Konzeption: Daniel Sirtes

Programmierung: Mathias Reichert

Prozessierung: Astrid Sohn

Bericht: Paul Donner

DZHW Berlin, 2019

Einführung

Funding Acknowledgements (FA) sind in wissenschaftlichen Veröffentlichungen enthaltene Informationen über die Förderung der der Veröffentlichung zugrunde liegende Forschung. FA-Daten sind in WoS ab 2008 verfügbar. Die in den Daten enthaltenen Angaben sind jedoch weitestgehend unstrukturiert und unbereinigt. So wurden beispielsweise über 12.000 Namensvarianten für die DFG in den Daten identifiziert (siehe auch Tabelle 1). Um die Daten für Analysen nutzbar zu machen, ist daher eine Bereinigung erforderlich. Ein vom iFQ/DZWH durchgeführtes Projekt hatte das Ziel, die in den WoS FA Daten enthaltenen Informationen über Förderorganisationen zu bereinigen und zu unifizieren, d.h. eine Vereinheitlichung der unterschiedlichen verwendeten Namensformen von Fördereinrichtungen war angestrebt. Nicht Gegenstand der Bereinigung waren dagegen die Angaben dazu, welche Wissenschaftler welche Förderungen erhalten haben und Angaben zu den Förderkennzeichen. Als Ausgangsdaten dienten die in den WoS-Daten enthaltenden Namensformen von Fördereinrichtungen, welche vom Anbieter automatisch aus den Fördertexten extrahiert werden, aber nicht weiter vereinheitlicht werden. Diese Daten sind in den KB Bibliometriedatenbanken in der Tabelle *fundingorganizations* enthalten. Daten zu Fördereinrichtungen, die nicht in WoS bzw. nicht korrekt aus dem Fördertext extrahiert werden konnten, gingen somit nicht in die Bereinigung mit ein.

Verfahren

Das angewendete automatische Verfahren basiert grundlegend auf dem hinreichend oft auftretendem gemeinsamen Vorhandensein von Abkürzungen und vollständigen Formen der Namen von Förderorganisationen im Datenfeld *fundingorganization* in den Ausgangsdaten. Abkürzungen von Einrichtungsnamen werden dabei automatisch über Mustersuche erkannt. Ausgehend von diesen Namensformen werden über Ähnlichkeitssuchen weitere Namensvarianten in einem mehrstufigen Prozess identifiziert. Als Datenquelle werden ausschließlich die Inhalte des *fundingorganization*-Feldes verwendet. Näheres siehe Referenzen [1-3]. Im Idealfall steht eine Abkürzung für eine Fördereinrichtung und alle auftretenden unterschiedlichen Namensformen sind dieser Abkürzung korrekt zugewiesen.

Ergebnisse

Die Ergebnisse des automatischen Verfahrens bestehen aus Tupeln von Abkürzungen, Namensformen und Ländercodes. Die Ergebnistabelle umfasst ca. 1 Mio. Einträge zu ca. 3600

Abkürzungen auf dem Datenstand von 2016. Zur Validierung wurden manuell Daten erhoben und mit den Ergebnissen des automatischen Verfahrens verglichen. Für 13 Förderorganisationen wurden alle Namensvarianten von Hand recherchiert. Mit den Daten von 9 dieser Einrichtungen kann sinnvoll die Vollständigkeit (*recall/true positive rate*) des automatischen Zuweisungsverfahrens abgeschätzt werden. Für die anderen fehlen die Daten für die automatische Bereinigung weil es keine gängige Abkürzung gibt (Studienstiftung), weil sie keine eigenständigen Förderorganisationen sind (Helmholtz-Gemeinschaft, Max-Planck-Gesellschaft) oder weil die Organisation aus nicht ersichtlichen Gründen in den automatisch bereinigten Daten nicht enthalten ist (Alexander von Humboldt-Stiftung). Diese Auswahl stellt somit keine valide Zufallsstichprobe dar und ist nicht repräsentativ. Die Ergebnisse des Vergleichs sind in Tabelle 1 dargestellt. Es zeigt sich, dass für 8 der Einrichtungen mit dem automatischen Verfahren sehr gute Ergebnisse hinsichtlich Vollständigkeit erzielt werden konnten. Unbefriedigend sind hingegen die Ergebnisse für die Volkswagen-Stiftung. Die Ursache liegt darin, dass in den Funding Acknowledgments in der Regel nicht die Kurzform VW sondern der vollständige Name verwendet wird. Dies belegen die Ergebnisse der manuellen Analyse. Durch das automatische Verfahren wurden zusätzliche Schreibweisen identifiziert, die nicht von Hand gefunden worden sind. Eine Ursache ist, dass die zugrunde liegenden Daten, auf die das automatische Verfahren angewendet wurde die Publikationen eines weiteren Publikationsjahres enthalten, da sie auf der Datengrundlage erstellt wurden, die ein Jahr später zur Verfügung stand als jene, die für die manuelle Recherche verwendet wurde. Es wurde keine statistische Auswertung dieser zusätzlich identifizierten Namensvarianten vorgenommen. Eine cursorsche Sichtung legt jedoch nahe, dass dies ganz überwiegend korrekt zugewiesene Namensvarianten sind.

Ergebnistabellen:

dzhwpdonner.fa_validation: manuell bereinigte Validierungsdaten für 12 Förderorganisationen

- abbr: Abkürzung des Namens der Förderorganisation
- fundingorganization: zugeordnete Namensvariante, entstammt der WoS-Tabelle fundingorganizations und kann mit dieser gejoint werden

dzhwpdonner.fa_automatic: automatisch bereinigte Daten

- abbr: Abkürzung des Namens der Förderorganisation
- fundingorganization: zugeordnete Namensvariante, entstammt der WoS-Tabelle fundingorganizations und kann mit dieser gejoint werden
- countrycode: Ländercode der Förderorganisation

Tabelle 1: Validierung des automatischen Verfahrens

Fördereinrichtung	Abkürzung	intellektuell erfasste Referenzdaten		von diesen durch automatisches Verfahren gefunden		ungewichtete Vollständigkeit	gewichtete Vollständigkeit	durch autom. Verfahren gefundene zusätzliche Namensformen
		verschiedene Namensformen	Namensformen	verschiedene Namensformen	Namensformen			
Deutscher Akademischer Austauschdienst	DAAD	2012	10978	2009	10973	0,999	1,000	1426
Fonds zur Förderung der Wissenschaftlichen Forschung	FWF	1660		1658		0,999		1019
Volkswagen-Stiftung	VW	343	3134	68	224	0,198	0,071	40
Fonds der Chemischen Industrie Deutsche	FCI	409	6283	409	6283	1,000	1,000	79
Forschungsgemeinschaft Bundesministerium für Bildung und Forschung	DFG	12755	125225	12697	125152	0,995	0,999	6056
Agence nationale de la recherche	BMBF	6863	34943	6850	34930	0,998	1,000	3233
Nederlandse Organisatie voor Wetenschappelijk Onderzoek	ANR	8455		8435		0,998		6014
Centro de Investigacion Biomedica en Red	NWO	4994		4985		0,998		3266
	CIBER	576		576		1,000		352

Einschränkungen

Das automatische Verfahren basiert auf einer rein musterbasierten Erkennung von Abkürzungen und damit in Verbindung stehenden Namensformen. Dies führt dazu, dass nur Einrichtungen, die genügend oft mit der Abkürzung und ausgeschriebenen Namensformen in den Daten auftreten, erkannt werden. Fördereinrichtungen, die üblicherweise mit vollem Namen genannt werden, wie beispielsweise die Volkswagen-Stiftung oder die Studienstiftung des deutschen Volkes, können mit dem Verfahren nicht behandelt werden.

In den Daten treten auch zahlreiche Abkürzungen auf, die nicht zu einer Förderorganisation gehören (z. B. die Unternehmensform KGaA). Es gibt Abkürzungen, die mehr als eine Einrichtung bezeichnen (z. B. ABC).

In den Daten existiert keine ID oder Ansetzungsform, die die Einrichtung eindeutig bestimmt. Es gibt lediglich die Abkürzung als Einstiegspunkt.

Die Bereinigung weiterer Daten ist aufgrund der sehr langen Laufzeiten des Programms nicht machbar und dadurch die Überführung in eine Standardkomponente der KB-Infrastruktur nicht realisierbar. Die Notwendigkeit für die Bereinigung neuerer Daten ist zudem auch geringer, weil inzwischen bereits bessere Datenquellen, insbesondere Crossref Funder Registry, existieren.

Verwendungsmöglichkeiten

Aufgrund der genannten Einschränkungen sollte von einer Verwendung der automatisch erzeugten Daten ohne weitere Validierungsaktivitäten abgesehen werden. Die jetzt vorliegenden (vor-)bereinigten Daten stellen jedoch zumindest eine geeignete Ausgangsbasis für Analysen dar. Erforderlich ist jedoch in jedem Fall eine weitere Prüfung der für die jeweiligen Analysezwecke benötigten spezifischen Daten, fallweise mittels manueller Stichprobenkontrolle.

Die Daten der manuell durchgeführten Bereinigung sind dagegen nutzbar, für sie gilt jedoch auch, dass sie nicht auf dem aktuellsten Stand sind, d.h. für aktuelle Jahrgänge sind ergänzende Bereinigungen vorzunehmen.

Literatur

[1] Sirtes, D.; Riechert, M.; Donner, P.; Aman, V.; Möller, T. (2015): [Funding Acknowledgements in der Web of Science Datenbank. Neue Methoden und Möglichkeiten der Analyse von Förderorganisationen](#), Studien zum deutschen Innovationssystem, Berlin: EFI.

[2] Sirtes, D.; Riechert, M., (2014): [A Fully Automated Method for the Unification of Funding Organizations in the Web of Knowledge](#). In: *Noyons, Ed (Ed.): Context counts: pathways to master big and little data. Proceedings of the 19th International Conference on Science and Technology Indicators, 2014 Leiden.* p. 594-597. DOI: 10.13140/2.1.3086.5285. ISBN 978-90-817527-1-8.

[3] Sirtes, D., (2013): Funding Acknowledgements for the German Research Foundation (DFG). The Dirty Data of the Web of Science Database and How to Clean It Up. in: Gorraiz, J. et al. (eds): Proceedings of the 14th International Society of Scientometrics and Informetrics Conference, Vienna: AIT GmbH, Volume 1, 784-795.