



kompetenzzentrum bibliometrie

Institutionenkodierung

Teilprojektbeschreibung

1. Problemstellung und Ziel

In den beiden interdisziplinären Literaturdatenbanken Web of Science (WoS) und Scopus sind pro Publikation zwar die institutionellen Adressen der Autoren erfasst, soweit diese im Original der jeweiligen Arbeit aufgeführt sind. Die Erfassung dieser Adressen lässt aber hinsichtlich Genauigkeit und Vollständigkeit sehr zu wünschen übrig. Dahinter verbergen sich verschiedene Probleme, für die es nicht in jedem Fall einfache Lösungen gibt.

Am bekanntesten ist das Problem der unterschiedlichen Schreibweisen („spelling variants“) von Adressen, dem die Datenbankhersteller seit geraumer Zeit durch Bemühungen zur (Teil-) Standardisierung der Adresseinträge entgegenwirken. Diesen Ansätzen sind jedoch enge Grenzen gesetzt, weil die Adressangaben schon in den Originalartikeln der wissenschaftlichen Zeitschriften nicht selten ungenau und unvollständig sind. Die Zeitschriften arbeiten mit unterschiedlichen (bzw. gar keinen) Vorgaben für die Autoren hinsichtlich der Nennung der relevanten Adressen. Entsprechend heterogen sieht die Praxis aus. Es gibt sogar Disziplinen, in denen die Nennung von abgekürzten Privatadressen statt der Institution nicht unüblich ist. Ein weiteres Problem sind die Adressangaben von Autoren, deren Arbeiten im Kontext von Gastaufenthalten an „fremden“ wissenschaftlichen Einrichtungen entstehen: in solchen Fällen werden zum Teil zwei Adressen für einen Autor angegeben, manchmal (aber nicht immer) mit entsprechender Qualifizierung.

Ein nicht zu unterschätzendes Problem stellen außerdem die Extremfälle von Publikationen dar, an denen Autoren von einer sehr großen Zahl von Institutionen beteiligt sind. Solche Fälle finden sich vor allem im Bereich der Hochenergiephysik und bei den großen klinischen Studien in der Medizin. Die an bestimmten Stellen zunehmende, zum Teil politisch ausdrücklich gewünschte Überlappung von Sektoren im deutschen Wissenschaftssystem (z.B. Max-Planck-Gruppen in Hochschulen, gemeinsame Berufungen bei HGF und Hochschulen, Hybrideinrichtungen wie KIT, neue Organisationsformen für Universitätsklinik wie Charité, UK-SH etc.) macht eine eindeutige Zuordnung von Publikationen zu Institutionen zusätzlich schwierig. Hinzu kommt, dass im Zeitverlauf die Institutionenlandschaft nicht stabil, sondern im Wandel ist.

Vor diesem Hintergrund wäre es unzulässig, valide bibliometrische Indikatoren institutionsbezogen auf der Basis der unbereinigten, in WoS oder Scopus vorgefundenen

Adressdatensätze generieren zu wollen. Dass dies in manchen Kontexten dennoch immer wieder versucht wird, trägt zur Gefahr des Missbrauchs von Bibliometrie im politischen Raum bei. Der unreflektierte Einsatz von kommerziellen, „einfach bedienbaren“ Standardprodukten wie etwa der „Essential Science Indicators“ (ESI) oder anderweitig publizierter bibliometrischer Rankings in Instanzen der Wissenschaftsverwaltung und -Politik birgt Risiken der Fehleinschätzung.

Das hier vorgeschlagene Teilprojekt zielt demgegenüber auf die Schaffung einer seriösen Basis für die bestmögliche Zuordnung der in WoS und Scopus erfassten Publikationen aller deutschen Institutionen.

2. Stand der Forschung

Die Probleme der Unifizierung und Kodierung von institutionellen Adressen für bibliometrische Analysen sind seit langem bekannt. De Bruin & Moed (1990) haben bereits vor 20 Jahren darüber berichtet. Dennoch gibt es bis heute relativ wenig Literatur, die sich dezidiert darauf bezieht. Bekannt ist, dass die Hersteller des WoS (bzw. des Vorläufers SCI) im eigenen Interesse zwar Anstrengungen zur Unifizierung und Standardisierung unternommen haben, dass diese aber an Grenzen stoßen, die ohne detaillierte Kenntnisse der regionalen und lokalen institutionellen Strukturen nicht zu überwinden sind. Eine Bereinigung der Adressen erforderte bisher letztlich immer auch manuelle Eingriffe und Korrekturen. Moed et al. haben einen Einblick in ihre Erfahrungen mit der Kodierung der niederländischen Adressen im SCI gegeben:

“The unification and classification of Dutch corporate source addresses in our database is almost complete now. Over 99% of the addresses has been unified and classified according to the system mentioned above. However, several of the addresses that have been unified and classified, are still 'problematic'. These institutes belong to overlapping categories, in which their status could not clearly be established or their identification (e.g., as a university department) was not 100% certain. Since these 'problematic' addresses in our database generally have a small output, we believe that we have unambiguously unified and classified the institutes responsible for well over 95% of the publications. This is sufficient to produce reliable results per institutional sector. Publications that could not be classified yet were not taken into account in the analyses.”
(Moed et al.1995, S. 393)

In der Zwischenzeit sind Adresskodierungen in zahlreichen bibliometrischen Einzelprojekten meist ad hoc und nur für den jeweiligen Einsatzzweck vorgenommen worden. Entsprechende Erfahrungen liegen vor allem in den Niederlanden und Belgien, Großbritannien, Australien und Skandinavien vor. Die Ergebnisse sind jedoch im Allgemeinen nicht öffentlich zugänglich.

Galvez & Moya-Anegón haben 2006 und 2007 über einen neuen Ansatz mit der Anwendung parametrisierter endlicher Graphen (FSG) bzw. Transduktoren (FST) berichtet, der an WoS-Daten entwickelt und später auch mit Daten aus Inspec, Medline, CAB Abstracts getestet wurde. Der Einsatz dieser NLP-basierten Methoden zur Standardisierung von Autorenadressen erscheint erfolgversprechend. Dennoch bleiben auch in diesem Fall absehbare Grenzen der automatisierten Behandlung, wie die Autoren selbst feststellen:

“It is important to point out that beyond the scope of the present work remain those problems originating in any errors or inconsistencies produced by abbreviations, transliteration differences, differences in spelling, or name changes. Nor do we tackle problems deriving from the absence in the address of the first institutional level, or difficulties in the assignment of each document to a center that may result from ambiguity or inconsistency in the use of different names to refer to a single institution, cases where a single same name may designate two or more separate institutions, or assigneeship reflecting different nationalities. The validation and correct institutional assignment of addresses is a task corresponding to experts.” (Galvez & Moya-Anegón 2007, S. 6)

Das gilt sicher umso mehr, wenn die Zuordnung von Personen (Autoren) zu Institutionen in bibliometrischen Analysen nicht starr, sondern flexibel gehandhabt werden soll. Dies wird im Kontext forschungspolitischer Anwendungen zunehmend nachgefragt, um z.B. auch virtuelle Einrichtungen wie institutionenübergreifende Projektverbände, Exzellenzcluster u.ä. angemessen analysieren zu können. Auch die Möglichkeit einer alternativen Betrachtung der Daten unter den Prinzipien von „work-done-at“ einerseits und “current potential” andererseits erhält zunehmend Bedeutung (vgl. Wissenschaftsrat 2008, S. 17). Die Datenbankhersteller haben darauf inzwischen mit der Einführung von personenbezogenen Kennungen reagiert (Thomson Reuters: <http://isiwebofknowledge.com/researcherid>). Angesichts des dadurch entstehenden enormen Kontrollpotentials regen sich in der scientific community allerdings Bedenken gegen derartige Systeme (Enserink 2009), so dass von einer flächendeckenden Umsetzung und Akzeptanz bisher keine Rede sein kann.

3. Projektbeschreibung

Um das Ziel der bestmöglichen Zuordnung der in WoS und Scopus erfassten Publikationen aller deutschen Institutionen zu erreichen, soll modular vorgegangen werden.

Zunächst sollen die Adressdatensätze eines aktuellen Datenbankjahrgangs (2008) des WoS soweit wie möglich mit automatischen Prozeduren kodiert werden. Dabei wird die Eignung der von Galvez & Moya-Anegón vorgeschlagenen Methoden getestet werden. Der Fokus liegt auf der möglichst vollständigen Kodierung aller deutschen Adressen auf der Ebene der jeweiligen Hauptinstitution (Hochschule, MPG-Institut, FhG-Einrichtung etc.). Ziel ist dabei, die Restmenge der ohne manuelle Nachbearbeitung nicht eindeutig identifizierbaren Adresseinträge mit vertretbarem Aufwand so klein wie möglich zu halten.

Im nächsten Schritt sollen die anhand des WoS-Jahrgangs gewonnenen Erfahrungen auf den Paralleljahrgang von Scopus übertragen werden. Soweit zu diesem Zeitpunkt bereits auf eine Matching-Tabelle zurückgegriffen werden kann, mit der die gesicherte wechselseitige Zuordnung der in beiden Datenbanken erfassten Dokumente ermöglicht wird, können entsprechende Verbesserungen am automatischen Kodierungsverfahren eingeführt werden, die den Rückgriff auf die jeweils relevanten Adresseinträge der anderen Datenbank nutzen.

Die nach Durchführung der automatischen Kodierung verbleibende Restmenge von Problemadressen, die nicht ohne weiteres als zu einer deutschen Hauptinstitution gehörig zu identifizieren sind, werden einer manuellen Sichtung und Bearbeitung un-

terzogen. Die dabei gewonnenen Erfahrungen, z.B. über typische Muster, sollen für die Einspeisung in spätere iterativ durchgeführte Kodierungsdurchläufe aufbereitet werden, soweit dies möglich und sinnvoll ist.

In einem weiteren Schritt werden jenseits einer bloßen Kodierung von Adressen nunmehr definitiv Publikationen zu (deutschen) Institutionen zugeordnet. Hierzu werden externe Quellen wie die Institutionendatenbank der DFG, Institutionenverzeichnisse von MPG, FhG, HGF und WGL etc. herangezogen. Ggf. kann zusätzlich auf Webseiten der Institutionen zurückgegriffen werden. Ziel an dieser Stelle ist, möglichst alle im Jahrgang 2008 der beiden Datenbanken enthaltenen deutschen Publikationen den passenden deutschen Hauptinstitutionen klar zuzuordnen. Soweit einschlägige öffentlich zugängliche Datenbestände zur Dokumentation des Publikationsoutputs bestehen (z.B. die Publikationsverzeichnisse von MPG und FhG), sollen diese genutzt werden.

Die detaillierte Zuordnung von Publikationen bis auf die unterste Ebene der jeweiligen Institution (z.B. Arbeitsgruppe, Lehrstuhl) ist exemplarisch für zwei Fälle vorgesehen, für die entsprechende Vorerfahrungen und Kenntnisse der Binnenstruktur vorliegen: die Ludwig-Maximilians-Universität (LMU) München und die Universität Bielefeld. Mit diesen beiden Fallstudien soll auch geklärt werden, wie gut die verschiedenen Quellen zur institutionellen Struktur (Adresseinträge aus WoS und Scopus, DFG-Institutionendatenbank, hausinterne Informationssysteme der betroffenen Institutionen, Selbstdarstellungen der Institute und Forscher im WWW) zur Deckung zu bringen sind. Ausgehend von den Datenbeständen des Jahrgangs 2008 soll ein Konzept entwickelt werden, um dem institutionellen Wandel über die Zeit Rechnung tragen zu können („Historisierung“ der institutionellen Kodierungen).

Auf der Basis der in den bisher erwähnten Schritten gewonnenen Erfahrungen werden dann die Prozeduren zur automatischen Kodierung auf möglichst alle verfügbaren Jahrgänge von WoS und Scopus ausgedehnt. Es wird zu prüfen sein, ob eine anschließende manuelle Nachkorrektur der deutschen Adressen in den anderen Jahrgängen sinnvoll und mit vertretbarem Aufwand durchzuführen ist.

Im letzten Schritt schließlich soll ein Verfahren entwickelt werden, um die von den Datenbankherstellern eintreffenden Lieferungen neuer Daten kontinuierlich kodieren zu können. Gemäß den in den übrigen Schritten gewonnenen Erkenntnissen sollen alle neu einlaufenden Adressdaten automatisch vorkodiert und die deutschen Publikationen so weit wie möglich dezidiert den passenden deutschen Institutionen zugeordnet werden.

Zusammengefasst besteht das Arbeitsprogramm aus sieben Modulen:

1. Vorkodierung eines aktuellen WoS-Datenbankjahrgangs (2008) mit automatischen Prozeduren. Test der von Galvez & Moya-Anegón vorgeschlagenen Methoden. Fokus auf deutschen Adressen
2. Anwendung der unter 1. gewonnenen Erfahrungen auf den gleichen Jahrgang von Scopus (2008)

3. Manuelle Nachkodierung der wichtigsten nach Durchführung der automatischen Kodierungen noch verbliebenen Problemadressen im ausgewählten Jahrgang der beiden Datenbanken
4. Zuordnung zu Hauptinstitutionen gemäß externer Quellen (DFG Institutionendatenbank, Institutionenverzeichnisse von MPG, FhG, HGF, WGL etc.)
5. Detailzuordnung für exemplarische Fälle (LMU, Uni Bielefeld), Entwicklung eines Konzepts zur Historisierung der Institutionenkodierung
6. Ausweitung der automatischen Kodierung auf die übrigen Jahrgänge
7. Etablierung eines Verfahrens zur kontinuierlichen Kodierung der neu einlaufenden Adressdaten

4. Erwartete Ergebnisse

Als Ergebnis des Teilprojekts soll eine mit automatischen Prozeduren herbeigeführte möglichst gute Standardisierung aller Adresseinträge aus WoS und Scopus bereitstehen und darüber hinaus eine weitgehend exakte Zuordnung aller deutschen Publikationen zu den relevanten deutschen Hauptinstitutionen. Die an den exemplarisch untersuchten Fällen vorgenommene Detailkodierung bis auf die untersten Hierarchieebenen soll Aufschluss darüber geben, inwieweit ein entsprechendes Vorgehen bundesweit flächendeckend realisierbar erscheint.

Institutionsbezogene bibliometrische Analysen sollen damit in Zukunft in Deutschland mit einem wesentlich höheren Qualitätsstandard durchgeführt werden können, als dies bei Verwendung der unbereinigten Ausgangsdaten von WoS und Scopus oder von Standardprodukten wie ESI möglich wäre. Für die Sicherstellung einer dauerhaft guten Zuordnung von Publikationen zu Institutionen würde damit im Kompetenzzentrum Bibliometrie die Voraussetzung geschaffen.

5. Literatur

De Bruin, R. E. & Moed, H. F. (1990), The unification of addresses in scientific publications. In: L. Egghe, R. Rousseau (Eds), *Informetrics 1989/90*. Elsevier Science Publishers, Amsterdam, pp. 65–78.

Enserink, M. (2009), Scientific Publishing: Are You Ready to Become a Number? *Science* 323, Nr. 5922 (27 March): 1662-1664.

Galvez, C. & Moya-Anegón, F. (2006), The unification of institutional addresses applying parametrized finite-state graphs (P-FSG). *Scientometrics* 69(2), 323-345.

Galvez, C. & Moya-Anegón, F. (2007), Standardizing formats of corporate source data. *Scientometrics*, 70(1), 3-26.

Moed, H., De Bruin, R. & Van Leeuwen, T. (1995), New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics*, 33(3), 381-422.

Wissenschaftsrat (2008), Bericht der Steuerungsgruppe zur Pilotstudie Forschungs-rating Chemie und Soziologie. Köln.