



kompetenzzentrum bibliometrie

## Entwicklung von Methoden und Indikatoren zur Bewertung von Fehlern in der Bibliometrie (Fehlerlehre)

Teilprojektbeschreibung

### 1. Problemstellung und Ziel

In der empirischen Sozialforschung übliche Methodenstandards und Fehlerlehren sowie darauf aufbauende Gütemaße sind in der Bibliometrie kaum entwickelt. Das ist nicht nur ein theoretisches Defizit, sondern zieht auch einen völlig unkritischen Umgang mit bibliometrischen Indikatoren, die nicht selten als „objektive Realitätsbeschreibungen“ betrachtet werden, nach sich.

Dieses Teilprojekt soll daher zunächst die in der Literatur bereits diskutierten und aus vorhandenen bibliometrischen Studien und Publikationserhebungen bekannten Fehler- und Biasquellen aufarbeiten, sodann ansatzweise eine theoretisch begründete Fehlerlehre formulieren und zugehörige Gütemaße bereitstellen sowie in einem weiteren Schritt Elemente einer empirisch unterfütterten Theorie des Zitierens entwickeln. Damit wird eine bisherige Lücke der Validierung und Qualitätssicherung der Indikatoren und Methoden geschlossen.

Die im Rahmen der Literaturanalyse identifizierten Methoden sollen systematisch mit den zur Verfügung stehenden Daten geprüft und die Unterschiede, die sich bei der Indikatorenkalkulation ergeben, ermittelt werden. Diese Ergebnisse dienen der Entwicklung von Gütemaßen für die Bewertung der bibliometrischen Indikatoren.

Was als „Fehler“ in bibliometrischen Analysen betrachtet werden kann, ist keineswegs selbstevident, denn bereits die Frage, was denn die Grundgesamtheit bibliometrischer Analysen sein soll, ist keineswegs trivial: Das, was Autoren als Publikationen führen? Das, was sich als fruchtbar erwiesen hat (also z.B. nicht die durchaus viel zitierten Fälschungen)? Das, was Datenbanken registrieren? Zumindest für Letzteres gibt es einen theoretisch fundierten Anspruch, nicht auf vollständige Erfassung von Publiziertem zu setzen, sondern „relevante“ Literatur zu selektieren (vgl. Bradford's Law). Inwieweit dies in den vorhandenen Datenbanken gelingt, ist allerdings nach wie vor eine offene Frage und zudem sehr disziplinabhängig.

In dem Maße, in dem bibliometrische Kennzahlen als Proxy für Reputation im Alltagsgeschäft der Wissenschaft genutzt werden, kommt einer Information über die „Belastbarkeit“ der ermittelten Indikatoren eine immer größere Bedeutung zu. Ein Ziel ist es daher, Gütemaße zu entwickeln, die je nach Einsatzgebiet auf die untersuchten Fachgebiete,

Institutionen oder Personen angewandt werden können und so eine Standardinformation für die interpretative Belastbarkeit der jeweiligen Kennzahl liefern. Eine zentrale Rolle wird die Fragestellung einnehmen, inwieweit sich die „Fehler“ und Unzulänglichkeiten wie etwa im Falle der verschiedenen Zählmethoden auf die Interpretation (etwa von Ratings, Rankings und Trends) auswirken und inwieweit sich die Interpretation bei Erhöhung respektive Senkung der Güte (wie etwa typisch bei Messwerten) nicht verändert.

Eine (disziplinspezifische) Methodologie und Fehlerlehre berührt einerseits sehr basale und theoretische Fragen, andererseits eher technische Konventionen und Vereinbarungen. Eine Fehlerlehre beinhaltet daher mehr als die Identifikation von „Fehlern“ in einem engen, technischen Sinne. Dieses Arbeitspaket verfolgt dementsprechend mehrere Ziele: a) die Entwicklung von Gütemaßen, die Nutzern bibliometrischer Kennzahlen eine Interpretationshilfe liefern und einer allzu naiven Verwendung von Kennzahlen vorbeugen, b) die begründete Wahl bestimmter Verfahren zur Verknüpfung von Merkmalen, die auch in anderen Arbeitspaketen zum Einsatz kommen (z.B. Match-Verfahren, die zitierendes und zitiertes Dokument verbinden), c) die Klärung der Indikatorqualitäten bibliometrischer Kennzahlen (was bildet sich in den bibliometrischen Daten ab? Welche Modifikationen sind zur Charakterisierung von Impact, Qualität, Resonanz, Aktivität etc. geboten?).

Das Arbeitspaket enthält daher auch verschiedene – unterschiedlich priorisierte – Module und ist auf einen längeren Bearbeitungszeitraum angelegt.

## **2. Stand der Forschung**

In der Fachliteratur der Wissenschaftsforschung werden schon längere Zeit und intensiv methodische Probleme, die sich im Kontext der Generierung bibliometrischer Indikatoren ergeben, diskutiert. Belastbares empirisches Wissen darüber, welche Auswirkungen die verschiedenen methodischen Vorgehensweisen auf die Indikatoren und deren Interpretierbarkeit bzw. Interpretation haben, gibt es hingegen kaum. Standards für die Konstruktion bibliometrischer Indikatoren und die Durchführung bibliometrischer Studien und Analysen gibt es ebenfalls nur bedingt (River Forrest 1996).

Auch mangelt es derzeit an einer systematischen und vollständigen Zusammenstellung der Probleme und deren Bewertung. Die technischen Fragestellungen sind nur teilweise in der Literatur erfasst und diskutiert. Großenteils handelt es sich um Erfahrungen aus bibliometrischen Studien, Machbarkeitsstudien und Anwendungen. Beispiele gibt es bezüglich der Verwendung von alternativen Matchkeys (Braun et al. 1985), aber auch hier sind dies eher erste Versuche als eine systematische Analyse.

## **3. Projektbeschreibung**

Ausgangspunkt dieses Projektes ist eine systematische Literaturanalyse, auf deren Basis eine Übersicht über die potenziell auftretenden Fehler und Interpretationsprobleme sowie deren Ausmaß und Auswirkungen entwickelt wird und Ansätze für Korrekturen und Gütemaße systematisiert werden. Die im Rahmen der Module entwickelten Gütemaße werden als Algorithmen in die bibliometrische Datenbank implementiert. Diese werden im Weiteren genutzt, um Indikatoren adäquat zu berechnen und deren Qualität und Aussagefähigkeit bewerten zu können. Parallel dazu sollen Algorithmen erstellt, ge-

prüft und implementiert werden, mit deren Hilfe bei der Konstruktion der bibliometrischen Indikatoren die identifizierten Fehler und ihr Ausmaß adäquat berücksichtigt werden können. Dazu wird eine Kooperation mit Spezialisten der Statistik und der Informatik angestrebt. Hierfür werden Mittel für einen Beratungsauftrag beantragt (s. Kostenaufstellung).

Aufbauend auf den Ergebnissen der einzelnen Module soll eine Theorie der Fehlerquellen und deren Auswirkungen auf verschiedenen Aggregationsstufen (Staaten, Institutionen, Forschungsgruppen und einzelne Wissenschaftler) entwickelt werden. Dabei sollen für die verschiedenen Faktoren auch der Charakter und Verzerrungsgrad abgeschätzt werden, um diejenigen Faktoren zu identifizieren, die besonders starke Einflüsse ausüben. Diese theoretischen Arbeiten werden begleitet von weiteren empirischen Untersuchungen, welche basierend auf den vorhandenen Datenbeständen einen Vergleich von empirischen und theoretischen Werte erlauben.

### *Modul 1: Zitations-Match-Verfahren und Fehler der Zitatraten*

Die Verfahren, die zur Ermittlung von Zitaten zum Einsatz kommen, werden zusammengetragen, verglichen und geprüft. Im Ergebnis wird ein geeignetes Verfahren für die Verknüpfung von zitiertem und zitierendem Dokument entwickelt. Dieses bildet die Voraussetzung für die korrekte Bestimmung von Zitatraten. Das Verfahren wird im Ergebnis in die Datenbank implementiert und stellt zugleich einen Input für die übrigen Teilprojekte dar.

Relevante Aspekte, die dabei behandelt werden sollen, betreffen in erster Linie die Matchkeys für die Verbindung von zitiertem und zitierendem Dokument sowie die Fehlerquellen der Zitatfindung. Welche Matchverfahren sind geeignet, um zitierende und zitierte Dokumente miteinander zu verbinden? Welche Auswirkung hat bspw. der Einsatz von umfangreichen „harten“ Matchkeys im Vergleich zu kurzen „weichen“ Matchkeys?

Bekannt ist außerdem, dass de facto alle in bibliometrischen Studien relevanten Datenbankeinträge wie Autorennamen, Adressen, Referenzen und Zeitschriftentitel häufig Schreibfehler aber auch eine große Zahl verschiedener Schreibweisen aufweisen.

Untersucht werden sollen daher die Fragen: Welche Auswirkungen haben fehlerhafte Referenzen (z.B. falscher Erstautor (Glänzel & Debakkere 2003) oder falsch geschriebene Namen und Vertauschung von Vor- und Nachnamen)?

Welche Auswirkungen haben die von den Datenbankbetreibern erzeugten Schreibfehler sowie die Behandlung von Umlauten und Sonderzeichen? Welche Konsequenzen ergeben sich aus der Verwendung verschiedener Namen oder bibliographischer Angaben (etwa Volume oder Issue) für Zeitschriften, (Braun et al. 1985). Ein Beispiel ist die Zeitschrift *Fortschritte auf dem Gebiet der Röntgenstrahlen und der bildgebenden Verfahren*, welche in Zitaten in der Regel als „ROFO“ bezeichnet wird, oder im Falle von *Nature* die Verwendung von Bezeichnungen der Issues in der Form „January“ bis „December“, aber alternativ auch 1 bis 12.

Häufig werden bibliographische Angaben in Referenzen verkürzt (Bsp. die Angabe „*Physical Review D*, Vol. 235“ wird verkürzt zu „PRD235“). Diese Verkürzungen beding-

ten früher eine Beschränkung der Matchkeys auf das erste Zeichen im Titel der Zeitschrift. Die Anfangsbuchstaben der Zeitschriften sind äußerst ungleich verteilt (I von International und J von Journal dominieren das Geschehen), was dazu führt, dass die Matchkeys sich auch oft als „zu weich“ und zu wenig trennscharf erweisen. Diese Fehler sind sowohl systematisch wie auch verzerrend (auf wenige Zeitschriften konzentriert). Es sollen daher die Fehler, die durch die Verwendung unterschiedlich verkürzter im Vergleich zu vollständigen Journaltiteln auftreten, untersucht und quantifiziert werden.

Das Ziel dieses Moduls ist die Entwicklung eines Gütemaßes, das durch Vergleich unterschiedlicher Match-Verfahren Hinweise auf potentielle Fehler gibt. Ein geeignetes Verfahren, das als Standard für das Kompetenzzentrum genutzt werden soll, wird damit identifiziert.

Die Arbeiten in diesem Modul sollen bereits 2009 beginnen und Ende 2009 abgeschlossen sein. Diese Arbeiten bilden eine Voraussetzung für andere beantragte Projekte (insbesondere das Projekt zur Ermittlung der erwarteten Zitatraten des Fraunhofer ISI).

### *Modul 2: Homonymen- und Schreibweisenproblematik von Autorennamen*

In Evaluationsstudien von Forschungsprogrammen und Ratings von Forschungsinstitutionen sind heute vermehrt Literatur- und Indikatorenrecherchen auf Basis von Autorennamen üblich und notwendig. Dies betrifft bspw. die Arbeiten im Rahmen des Ratings des Wissenschaftsrates aber auch die Evaluation konkreter Förderprogramme, in denen die Wirkungen dieser Programme untersucht werden sollen (z.B. des Emmy Noether Nachwuchsrgruppenleiter-Programms der DFG). In letzter Zeit immer wieder diskutiert wird die Erstellung von Autorennamendatenbanken bzw. die Einführung von Researcher IDs, die eine bessere Zuordnung von Publikationen zu Personen ermöglichen sollen.

Derartige Arbeiten an eindeutigen Autorenidentifikatoren gehen zurück auf zwei zusammenhängende Probleme. Einerseits gibt es eine Reihe von häufig auftretenden Namen, Homonymen, die dazu führen, dass Publikationen nicht eindeutig einer Person zugeordnet werden können. Andererseits treten für gleiche Namen häufig unterschiedliche Schreibweisen auf. Unterschiedliche Schreibweisen entstehen zum Beispiel durch das Vorhandensein mehrerer Vornamen, Schreibweisen von Umlauten, aber auch Akzenten und weiteren Nicht-ASCII Zeichen in ausländischen Autorennamen. Da heute in Datenbanken mehr und mehr auch größere Zeichensätze als das frühere sehr beschränkte ASCII Verwendung finden, nehmen diese Schreibweisen zu. Weiter führen auch Namensänderungen etwa bei Heirat zu mehreren Schreibweisen. Besondere Anforderungen an die Homonymtrennung stellen Geschwister und Ehepaare, welche in demselben Fachgebiet arbeiten – zum Beispiel Cole S. und Cole J.R. oder Marie Curie und Pierre Curie.

Zur Behandlung dieser Probleme und Fehlerquellen werden von verschiedener Seite Anstrengungen unternommen. Bspw. hat der Datenbankbetreiber Thomson Reuters sog. Researcher IDs eingeführt. Auf freiwilliger Basis können Autoren Angaben zur Person und ihren Publikationen machen und pflegen, so dass die korrekte Zuordnung von Personen und Autoren erleichtert wird. Anstrengungen in Richtung einer vollständigen Autorennamendatenbank gibt es auch von der Deutschen Nationalbibliothek.

Zur Homonymtrennung und Unifizierung von Schreibweisen werden zurzeit eher manuelle Verfahren eingesetzt, welche auf dem Vergleich bibliographischer Angaben basieren. So geben etwa ähnliche Koautorenlisten, aber auch gleiche Autorenadressen, gleiche Zeitschriften und Forschungsfelder Hinweise für die Trennung bzw. Zusammenführung. Die bisherigen Verfahren zeichnen sich jedoch durch drei Probleme aus. Erstens muss für jede bibliometrische Studie wieder neu mit dem zeitaufwändigen Trennen und zusammenführen begonnen werden; die Erfahrung aus früheren Studien ging mangels der Möglichkeit zur Implementation der Informationen in Datenbanken verloren. Zweitens sind die derzeitigen Verfahren nicht perfekt und weisen etwa für stark interdisziplinäre Gebiete große Fehlerquellen und Mängel auf; so werden etwa statistische Anwendungen in den Lebenswissenschaften in Zeitschriften der Lebenswissenschaften publiziert und werden dann nicht mehr den Autoren aus der Statistik zugeordnet. Drittens verbot sich im manuellen Betrieb der stark rechnerintensive Einsatz von etwa semantischen Verfahren, welche auf Text- und Stilvergleichen basieren.

Diese Probleme sollen in diesem Modul auf mehreren Ebenen angegangen werden: Zunächst sollen die Fehlerquellen und neuere Lösungsansätze zusammengestellt und verglichen werden, um Kenntnisse der Auswirkungen und Verzerrungen durch die Mängel und Fehler zu generieren. Gütemaße sollen in Abhängigkeit von den verschiedenen Namen und Schreibweisen eine Bewertung der potenziellen Fehler liefern und in die Datenbank implementiert werden. Anschließend soll unter Nutzung der Gütemaße die Auswirkungen der Fehler auf Zitatraten oder Publikationszahlen untersucht werden.

Darüber hinaus sollen bestehende Verfahren zur Behandlung von Homonymen und Variationen in Schreibweisen automatisiert werden und durch semantische Verfahren ergänzt werden. Die Einsatzmöglichkeiten vorhandener semantischer Programme, die bspw. zur Erkennung von Plagiaten aufgrund von ähnlichen Textmustern und -sequenzen eingesetzt werden (zum Beispiel das Programm TURNITIN<sup>®</sup>) sollen geprüft werden.

Zurückgegriffen werden soll auch auf die bereits von den Datenbankherstellern begonnenen Anstrengungen bei der Zuordnung von Autorennamen und Adressen. Außerdem lassen sich die Informationen aus den Publikationen mit nur einem Autor und dessen Adressangaben nutzen, um die Verfahren zu verfeinern. Die in diesem Modul gewonnenen Informationen lassen sich weiterhin als Input für die Adresscodierung und die korrekten Zuordnung von Autoren und Adressen nutzen und bieten somit Anknüpfungsmöglichkeiten zum Projekt des IWT.

### *Modul 3: Zitate als Impactschätzung*

Abgesehen von den im ersten Modul beschriebenen Fehlern, die sich aus den Zitate-Matching-Verfahren ergeben, sind Zitationsindikatoren von weiteren Fehlerquellen betroffen. Nicht bekannt sind die Art der Auswirkung, die diese hervorrufen, und deren Ausmaß.

Disziplinen unterscheiden sich hinsichtlich des Zitierverhaltens, wie auch hinsichtlich des Publikationsverhaltens. Unterschiede gibt es bspw. bezogen auf den Zeitpunkt des Erreichens des „Höhepunkts der Zitationsaktivität“. Disziplinen unterscheiden sich somit bezogen auf die Zitatehalbwertszeiten (Anzahl Jahre, bis die Hälfte der

Zitate erreicht ist). Hieraus ergeben sich Konsequenzen hinsichtlich der Frage welches der geeignete Zeitraum ist, der bei Zitationsanalysen zu berücksichtigen ist. Der Anspruch auf Aktualität bei der Bewertung von Programmen und Institutionen verleitet zur Verwendung so genannter „Kurzzeitimpacts“, die jedoch zu hohen Verzerrungs- und Fehlerraten führen können.

Im Rahmen dieses Moduls soll disziplinenpezifisch die Eignung der verschiedenen Zitatfenster geprüft werden und ein Gütemaß entwickelt werden, dass die Qualität der Zitatindikatoren in Abhängigkeit vom verwendeten Zitatfenster bewertet. Wiederum werden diese Gütemaße in die Datenbank implementiert und für die Indikatorenberechnung verfügbar gemacht. In der Folge werden Vergleiche von Resultaten unter Verwendung verschiedener Zitatfenster vorgenommen.

Neben dieser Problematik der zeitlichen Zitatekonzentration tritt auch das Problem der Zitatekonzentration auf oftmals wenige Artikel. Hierdurch wird die Bildung geeigneter Indikatoren für Aggregate (zum Beispiel Zeitschriften, Forschungsgruppen, Institutionen) mit den herkömmlichen Mittelwerten und Raten erschwert bzw. es ergeben sich stark verzerrte und kaum interpretierbare Ergebnisse. Mögliche Ursachen dieses Problems werden in Mertons Arbeiten zum „Matthäus-Effekt“ diskutiert und das Ausmaß und die Modellierung dieser Verzerrung wurde von Price 1978 in seiner Arbeit über „Cumulative Advantage“ angegeben. Weitgehend unbekannt sind nach wie vor die wirklichen Verzerrungen und die Konsequenzen für die Indikatorenkonstruktion.

Ein weiteres Problem ergibt sich im Kontext der Selbst- oder Eigenzitate. Zu klären ist die Frage, ob diese bei Impactanalysen verwendet oder ausgeschlossen werden sollten und welche Konsequenzen sich aus den verschiedenen Lösungsmöglichkeiten ergeben. Dienen Zitate eher Quellenangaben früherer Resultate, so sind Selbstzitate einzuschließen. Im Sinne einer Impactschätzung müssten diese aber eigentlich ausgeschlossen werden. In diesem Modul sollen die Auswirkungen des Einschlusses bzw. des Ausschlusses der Selbstzitate auf die Zitationsindikatoren untersucht werden.

Das Ziel ist einerseits, eine vollständige Information über die Auswirkungen des Ein- oder Ausschlusses von Selbstzitationen durch vollständige Rechnung der verschiedenen Varianten zu erhalten. Tabellen mit Zitatangaben nach „citing year“ sind mit relationalen Datenbanken möglich und gestatten eine explizite und synoptische Darstellung der Resultate simultan für alle Zeitfenster. Damit entsteht ein implementierbares und jährlich erweiterbares explizites Gütemaß. Zur Problematik der Zitatekonzentration auf wenige Artikel soll ein Algorithmus eines geeigneten Konzentrationsmaßes (z.B. Gini-Index oder Gini-Grafik) ausgewählt und implementiert werden. Im Gegensatz zu den hart implementierbaren Zitatraten muss dieser Algorithmus „on the fly“ einsetzbar sein und die Konzentration und weitere Indikatoren zu den jeweils abgerufenen Datensätzen berechnen.

Ziele dieses Moduls sind somit die Entwicklung und Implementierung von Gütemaßen für die Beurteilung der disziplinenpezifischen Eignung der verschiedenen Zitatfenster und die Beantwortung der Frage nach der Verwendung von Selbstzitationen und den damit verbundenen Wirkungen auf die Zitationsindikatoren. Auch hier werden entsprechende Gütemaße ermittelt und implementiert. Darüber hinaus wird ein Algo-

rithmus entwickelt, um die Zitatkonzentration zu bewerten und somit die Güte aggregierter Zitzahlen zu bewerten.

#### *Modul 4: Coauthorschaften und Position in der Autorenreihe, Institutionelle Adressen*

Publikationen sind zunehmend nicht das Werk eines einzelnen Autors. Häufig stellt sich die Frage nach dem Beitrag, den der Einzelne zu einer Publikation (und damit zu der damit verbundenen Forschung) geleistet hat. Ist die Reihenfolge der Autoren ein Maß für den Beitrag des Einzelnen? In den verschiedenen Disziplinen haben sich unterschiedliche Gewohnheiten bei der Festlegung der Autorenreihenfolge entwickelt. So wird in einigen Disziplinen Erstautorschaft mit Hauptautorenschaft gleichgesetzt, während in anderen Gebieten Letztautorschaft auf die Führung und Leitung der Forschungsarbeiten hinweisen soll und Projektmitarbeitende zu Anfang der Autorenliste stehen. In anderen Disziplinen werden die Autoren einfach alphabetisch gelistet.

Ein Problem ergibt sich, wenn Publikationen, wie heute gebräuchlich, auch Forschungsprogrammen, Staaten, finanzierenden Institutionen, durchführenden Institutionen oder auch Forschungsgruppen oder Einzelforschern zugeordnet werden sollen. Ist es adäquat, jedem die Publikation vollständig zuzuordnen, oder ist eine Teilzuordnung adäquat? Wenn letzteres der Fall ist, welcher Anteil ist korrekt? Welche Unterschiede ergeben sich aus der unterschiedlichen Positionierung unter den Autoren? Diesen Fragen wird in diesem Modul nachgegangen, um die Frage nach den Auswirkungen und Verzerrungen zu klären, die sich durch die verschiedenen Zuordnungen und Zurechnungen des Beitrags einzelner Autoren ergeben (bzw. darauf aufbauend für Institutionen, Länder etc.). Welche Zählmethoden sind geeignet, die Beiträge adäquat zu reflektieren? Die Auswirkungen und deren Konsequenzen sollen analysiert, quantifiziert und synoptisch dargestellt werden. Dazu werden Algorithmen erstellt, welche simultan für sämtliche Zuordnungsvarianten und Zählmethoden die entsprechenden Indikatorwerte rechnen. Im Bereich der Zählmethoden (full count versus fractional count) sind diese Algorithmen teilweise bereits entwickelt und publiziert worden (vgl. Gauffriau et al. 2007 für eine Zusammenstellung dieser Algorithmen) und können genutzt werden.

Zu den eher theoretischen Fragestellungen, die in diesem Modul behandelt werden sollen, gehört aber neben der expliziten Zuordnung auch die Frage nach der sinnvollen und angemessenen Zuordnung. Ist es angemessen, Publikationen den Autoren zuzuordnen oder den Autorenadressen, welche die beherbergenden Institutionen symbolisieren, oder bewirkt die unterschiedliche Zuordnung überhaupt eine Veränderung der Resultate oder deren Interpretation? Mit den Algorithmen zur synoptischen Darstellung ergibt sich auch hier eine weitere Grundlage für eine spätere Methodenwahl oder Standardbestimmung für Studien und Anwendungen der Bibliometrie.

Die Ergebnisse dieses Moduls können auch dazu dienen zu entscheiden, welche Teile der aufwändigen Codierungs- und Homonymtrennungsverfahren prioritär resp. von geringerer Auswirkung und damit Bedeutung sind.

#### *Modul 5: Gütemaße der Abdeckung*

Die großen Literaturdatenbanken unterscheiden sich zum Teil deutlich hinsichtlich ihres Inhalts. Die Hersteller nutzen unterschiedliche Kriterien zur Auswahl der zu erfassenden Literatur. In einigen Datenbanken stehen fachliche Abgrenzungen im Vordergrund, während andere auf eine möglichst große Breite und Multidisziplinarität

setzen. Unterschiede gibt es auch hinsichtlich der Daten, die für die einzelne Artikel erfasst werden (nur bibliographische Angaben oder aber auch Referenzen und Zitate) Für bibliometrische Studien im evaluativen Kontext stellt sich somit die Frage welches die geeigneten Datenbanken sind und welche Auswirkungen die unterschiedlichen Inhalte auf die bibliometrischen Indikatoren haben.

Das Publikationsverhalten unterscheidet sich zwischen den Disziplinen. Dazu gehört auch, dass unterschiedliche Kommunikationsmedien verwendet werden, die in der Datenbank wiederum unterschiedlich abgedeckt werden. So kommt bspw. in den Geistes- und Sozialwissenschaften Monografien eine große Bedeutung zu, während Proceedings in den Ingenieurwissenschaften und der Informatik wichtig sind. Hinzu kommt, dass Disziplinen unterschiedlich stark national oder international ausgerichtet sind, was sich nicht zuletzt in der Sprache der Publikationen niederschlägt.

Zwei Teilprobleme sollen in diesem Modul adressiert werden. Zum einen das eher theoretische Problem, was überhaupt eine angemessene Abdeckung des Outputs einer Disziplin ist und ob dieser wiederum adäquat in den Datenbanken erfasst wird. Die Frage, die sich stellt ist: Soll das als Literatur erfasst werden, was von Wissenschaftlern als wissenschaftliche Literatur wahrgenommen oder deklariert wird? Wiederholt ist darauf hingewiesen worden, dass einerseits die Erfassung der Literatur äußerst unterschiedlich und teilweise nach nicht nachvollziehbaren Kriterien erfolgt. Daraus ergeben sich Verzerrungen und Fehler, welche aber in ihrem Ausmaß und ihrer Bedeutung für die Interpretierbarkeit und Interpretation bei der Bewertung unter Verwendung bibliometrischer Indikatoren wenig bis gar nicht bekannt sind. Im Rahmen dieses Moduls sollen vergleichend für die beiden Datenbanken Web of Science und Scopus, zu einem späteren Zeitpunkt auch vergleichend für weitere Datenbanken, die Unterschiede und deren Auswirkungen auf die bibliometrischen Indikatoren, die sich aus unterschiedlichen Abdeckungsgraden ergeben, ermittelt werden. Daraus ergibt sich eine Grundlage für eine Bewertung der Adäquanz der erfassten Literatur in der Datenbasis für künftige bibliometrische Studien.

Untersucht werden sollen daher die Effekte, die die Auswahl der Datenbank auf die zitatenbasierten Kennwerte (Scopus versus WoS) hat. Neben der rein inhaltlichen Abdeckung werden die Auswirkungen der folgenden Faktoren untersucht: (1) Die Berücksichtigung der verschiedenen Publikationstypen als zitierendes und als zitiertes Dokument und deren Auswirkungen auf die Zitatmaße. Welche Dokumenttypen sind in der Konsequenz bei der Berechnung zu berücksichtigen? (2) Die Berücksichtigung so genannter „non source items“ aus den Referenzen der erfassten Artikel und deren Wirkungen auf die Zitationsindikatoren. Diese Frage ist zudem relevant für die Beantwortung der Frage nach der Adäquanz der Abdeckung der Forschung durch die verschiedenen Datenbanken bzw. der Notwendigkeit einer Erweiterung der teilweise ungenügend abdeckenden Datenbasis.

### *Modul 6: Gütemaße für Vergleichsgruppenbildung und Normierungsverfahren*

Wie in vielen Bereichen der Indikatorik weisen auch bibliometrische Indikatoren zunächst keine absolute Skala auf. Diese Feststellung geht zurück auf die Begutachtung einer der frühesten Anwendungen der Indikatoren, den ersten „Science and Engineering Indicators“ 1972 der National Science Foundation. Notwendig sind somit geeignete Vergleichsgruppen zur Bestimmung relativer Werte. Während in anderen Wissenschaften, etwa den Sozialwissenschaften, den Lebenswissenschaften, der Psychologie und der

klinischen Medizin, eine umfangreiche Literatur zu Versuchsplanung und Referenzgruppenbildung vorhanden ist und die Auswirkung von verschiedenen Methoden zur Vergleichsgruppenbildung und ihren Vor- und Nachteilen eingehend untersucht worden sind, gibt es hierzu in der Bibliometrie gerade erste Ansätze. Dieser Mangel wirkt umso schwerer, als es in der Forschung eine bedeutende Herausforderung ist zu bestimmen, was überhaupt vergleichbar ist. Sind etwa Publikationen in verschiedenen Disziplinen vergleichbar oder sind Zitate über Disziplinen hinweg vergleichbar? Sind Zitate der verschiedenen Publikationen in Zeitschriften und anderen Aggregaten vergleichbar? (s. dazu auch Modul 7 zur Heterogenität von Aggregaten). Sollen Forschende mit ihren Kooperationspartnern verglichen werden oder sollen Publikationen mit ähnlichen Referenzlisten zu Vergleichszwecken herangezogen werden? Sollen Zitationen an den vergebenen Referenzen gemessen werden?

Das Ziel dieses Moduls soll somit die Zusammenstellung der Methoden zur Referenzgruppenbildung, eine kritische Würdigung deren Eignung in verschiedenen evaluativen Kontexten und ein expliziter Vergleich der Resultate der verschiedenen Methoden (Gütemaß) sein.

Dazu wird zunächst die vorhandene Literatur zusammengestellt und ausgewertet. Herangezogen wird einerseits die bibliometrische Literatur zu Ansätzen bibliometriespezifischer Referenzgruppenbildung, andererseits aber auch die Literatur zu Referenzgruppen und Versuchsplanung im Allgemeinen. In einem weiteren Schritt sollen die verschiedenen Kontexte der Anwendung bibliometrischer Indikatoren charakterisiert und auf den Einsatz von Referenzgruppen untersucht werden. So sind etwa für die Ratings des Wissenschaftsrats andere Referenzgruppen adäquat als etwa für Leistungsorientierte Mittelvergaben in der klinischen und vorklinischen Medizin. In diesem Schritt soll auch auf die bestehenden Erfahrungen aus Studien und Evaluationsstudien der Konsortialpartner zurückgegriffen werden. In einem dritten Schritt werden wiederum durch expliziten Vergleich die Auswirkungen der verschiedenen Vorgehensweisen studiert (Gütemaßentwicklung). Die resultierenden Empfehlungen und Vorgehensweisen sowie ihre Vor- und Nachteile in den verschiedenen evaluativen Kontexten sollen in einem Handbuch oder in anderer geeigneter Form publiziert werden.

Dieser Modul hat starke Bezüge zum Teilprojekt „Erwartete Zitatraten und Klassifikationen“ des Fraunhofer ISI, da diese Klassifikationen eine der Methoden zur Gewinnung von Referenzgruppen darstellen. Die Berechnung und Implementierung erwarteter Zitatraten und die anschließende Normierung ist dann der Vergleich mit der Referenzgruppe in den Fällen, in denen keine „natürliche“ Referenzgruppe (etwa „vor Programmbeginn“, „während Programm“ und „nach Programmende“ sind im Falle der Evaluation von Forschungsprogrammen sicher sinnvolle und intuitive Referenzgruppen) vorliegt. Starke Bezüge gibt es auch zum Teilprojekt „Institutionenkodierung“ des IWT Bielefeld, da auch in diesen Projekten Referenzgruppen gebildet werden.

#### *Modul 7: Nutzung von Journalcharakteristika zur zeitnahen Publikationsbewertung*

Mit Zitationsindikatoren kann aufgrund der unterschiedlichen Zitationsgewohnheiten eine Bewertung der Publikationen oft erst nach längerer Zeit vorgenommen werden. In vielen evaluativen Kontexten sind aber auch zeitnahe Bewertungen erforderlich, um etwa bereits im Frühstadium von Programmen korrigierend eingreifen zu können. An-

dererseits ist eine Bewertung (wie bereits in Modul 3 „Zitate als Impactschätzung“ und Modul 5 „Gütemaße der Abdeckung“) und angemessene Erfassung unumgänglich.

Zum Zweck der zeitnahen Bewertung der Publikationsstrategie wurden bislang Journal Impact Factors (Mittlere Zitationsrate der Publikationen eines Zeitschriftenjahrgangs, JIF) errechnet und im Journal Citation Report veröffentlicht. Neben der grundsätzlichen Frage, ob einzelne Publikationen angemessen mit Zeitschriftenmittelwerten bewertet werden können, weisen diese Mittelwerte eine Reihe von Problemen auf. Ein Vergleich der Zitatverteilung in den Zeitschriften *Physical Review D* und *Lancet* ergab z.B., dass die Zitatverteilung im *Lancet* wesentlich stärker auf einige sehr hoch zitierte Publikationen konzentriert ist, während in *Physical Review D* zwar relativ geringe Spitzenwerte (fast acht mal kleiner), aber dafür eine sehr ausgeglichene Verteilung aufzufinden ist (lediglich ca. 5% unzitierte Publikationen stehen ca. 37% Unzitierten im *Lancet* gegenüber). Auch ein Vergleich der Mediane der Zitatverteilung ergab ein ähnliches Resultat, dass mit einem Medianwert von 7 die ca. 5000 Artikel aus *Physical Review D* wesentlich ausgeglichener zitiert werden als die ca. 1700 Artikel im *Lancet* mit einem Medianwert von zwei Zitaten. Im Gegensatz zu diesen parallelen Resultaten robuster Indikatoren (Integrale der Konzentrationen nach Gini und Lorentz, sowie Mediane sind äußerst robuste Maße) zeigen die Mittelwerte in die umgekehrte Richtung: Der JIF im *Lancet* liegt mit ca. 27 Zitaten pro Publikation deutlich höher als derjenige in der *Physical Review* mit ca. 12 Zitaten pro Publikation.

Somit stellt sich die Frage nach der grundsätzlichen Eignung der Aggregatwerte bzw. nach deren adäquater Berechnung. Ein Ziel muss ein Maß sein, das sich durch robuste Charakteristika auszeichnet, d.h. auch die Robustheit dient als Charakteristikum zur Bewertung der Güte des Maßes. Entwickelt werden soll ein Maß zur angemessenen und robusten Bewertung des Impact (oder Wirkungsmaße) und der Bewertung dessen Güte auf Basis expliziter Vergleiche. Die Konstruktion des Maßes soll anhand der Zeitschriften erfolgen, da in allen anderen Aggregaten zusätzliche verzerrende Faktoren wie Zuordnung, Zählmethoden oder Kodierungen die Verteilungen verfälschen können. Insbesondere sollen auch Verfahren der nicht-parametrischen Statistik wie etwa die sehr robusten Quantilswerte (etwa ergeben einfache Quantilswerte gemessen am Maximalwert robuste Maße der Konzentration resp. Ausgeglichenheit) genutzt werden. Zu untersuchen sind u.a. vergleichende grafische Darstellungen der gesamten Verteilungen wie etwa absteigend sortierte Zitate gemessen am Maximalwert der Zitate. Zu klären ist, ob auch diese Verteilungen von der Länge der Zitatfenster abhängen.

Damit entsteht einerseits eine Reihe von robusten Maßen für zeitnahe Bewertung von Zeitschriften aber auch der Publikationsstrategie und des Impacts.

### *Modul 8: Schritte in Richtung einer Theorie der Fehlerquellen*

Oftmals beklagt wird das Fehlen einer „Theorie des Zitierens“. Versuche, diese zu entwickeln, gab es in der Frühphase der Bibliometrie (vgl. Moravcik & Murugesan 1975; Chubin & Moitra 1975). Weitere Schritte in dieser Richtung wurden dann von Cronin (1984) unternommen. Das Diskussionspapier von Leydesdorff (1998) löste zwar eine starke Debatte in der Zeitschrift *Scientometrics* aus, konnte aber noch keine vollständige Klärung dieser für eine angemessene Interpretation zentralen Frage bringen. Die Frage nach der Interpretation der Zitationsindikatoren wurde dann weiter in einer Literaturstudie von Bornmann und Daniel (2008) diskutiert.

Diesem Desideratum der Wissenschaftsforschung wollen wir uns auf der Basis der Analyse eines Zufallssamples aus WoS und Scopus nähern. Dabei sollen die Autoren zu den Referenzen jeweils eines ihrer Artikel befragt werden und insbesondere Auskunft geben über Motiv, Kenntnis der Quelle und die Funktion des Zitats im eigenen Artikel. In gleicher Weise wollen wir die Frage der „Autorschaft“ behandeln. Aktuell gibt es nicht nur Hinweise darauf, dass die Verleger verstärkt Einfluss auf die Referenzen nehmen, um den Impact der jeweiligen Zeitschrift zu steigern, sondern auch darauf, dass insgesamt die Funktion des Zitates immer stärker durch strategische Momente bestimmt wird.

Dieser – über die reine Datenanalyse hinausgehende – Schritt, erscheint uns besonders wichtig, weil bibliometrische Kennzahlen selbst zum Bestandteil der Governance und des Managements der Forschung geworden sind (vgl. zum Beispiel die auf bibliometrischen Indikatoren basierende Leistungsorientierte Mittelvergabe (LOM)). Im technischen Sinne wird es damit fraglich, inwieweit bibliometrische Messungen tatsächlich „nicht reaktiv“ sind.

#### **4. Erwartete Ergebnisse**

Im ersten Modul entsteht ein Matchverfahren für die Zuordnung von Zitaten zu den zitierenden Publikationen. Dieses wird in die Datenbank implementiert und bildet dann die Grundlage für die weiteren Datenerhebungen und Indikatorenberechnungen.

Darüber hinaus werden Fehler und Gütemaße für bibliometrische Indikatoren auf unterschiedlichen Aggregationsstufen (bspw. Personen, Institutionen) ermittelt und in die Datenbank implementiert. Die entsprechenden Algorithmen zu deren Berechnung werden ebenfalls in die Datenbank implementiert. Somit wird eine dauerhafte Qualitätsprüfung der Daten ermöglicht und ein Beitrag zur Qualitätssicherung und -prüfung geleistet. Die Berechnung der Gütemaße wird periodisch aktualisiert.

Mit diesen Fehlerlehren werden auch Aussagen über die Machbarkeit von und die Voraussetzungen für bibliometrische Studien in den verschiedenen Wissenschaftsgebieten möglich sowie Standards für deren Durchführung bestimmt.

Weitere Ergebnisse sind Berechnungsalgorithmen für die geprüften Standardindikatoren und Basisindikatoren. Diese werden nach Erstellung und Prüfung direkt in die Datenbank implementiert. Die Methodologie wird für die Benutzer der Datenbank und Indikatoren in einem Handbuch zusammengestellt.

Die Projektergebnisse werden in ausgewiesenen Fachzeitschriften publiziert (bspw. *Scientometrics*) und auf Fachtagungen der Scientific Community (z. B. der Conference of the International Society for Scientometrics and Informetrics ISSI 2011) präsentiert und diskutiert. Hierzu wird eine fünftägige Ausland-Dienstreise beantragt.

Darüber hinaus erwarten wir, dass sich die Ergebnisse in künftigen Anwendungen niederschlagen, die entwickelten Methoden in den künftig durchzuführenden bibliometrischen Studien verwendet werden und so die Qualität und Aussagefähigkeit dieser erhöht wird.

## 5. Literatur

Anderson, J. P.; Collins, M. D.; Irvine, J.; Isard, P. A.; Martin, B. R. & Narin, F. (1988): "On-Line Approaches to Measuring National Scientific Output--A Cautionary Tale", *Science and Public Policy* 15, 153-161.

Bornmann L. & Daniel H.-D. (2008): "What do citation counts measure? A review of studies on citing behavior", *Journal of Documentation* 64, 45-80.

Braun, T.; Glänzel W. & Schubert A. (1985): *Scientometric Indicators*, World Scientific.

Chubin B. & Moitra S. (1975): "Content analysis of references: Adjunct of alternative to citation counting?", *Social Studies of Science* 5, 423-441.

Cronin B. (1984) "The Citation Process: The Role and Significance of Citations in Scientific Communication", Taylor Graham.

Gauffriau, M.; Larsen, P. O.; Maye, I.; Roulin-Perriard, A. & von Ins, M. (2007): "Publication, cooperation and productivity measures in scientific research", *Scientometrics* 73: 175.

Glänzel W. & Debackere K. (2003): "On the opportunities and limitations in using bibliometric indicators in a policy relevant context", in: *Bibliometric analysis in science and research: Applications, benefits and limitations, 2nd Conference of the Central Library*, Jülich, 5.-7. November 2003, Forschungszentrum Jülich.

Leydesdorff L. (1998): "Theories of citation?", *Scientometrics* 43 5-25.

Moravcik, M.J. & Murugesan P. (1975): "Some results on functions and quality of citations" *Social Studies of Science* 5, 6-92.

Price D.J. de Solla (1963): *Little Science Big Science*, Suhrkamp.

*Scientometrics* 35 (1996), "River Forrest Workshop" und Beiträge darin.