

Joachim Griesbaum, Thomas Mandl,
Christa Womser-Hacker (Hrsg.)

Information und Wissen: global, sozial und frei?

Proceedings des 12. Internationalen Symposiums
für Informationswissenschaft (ISI 2011)

Hildesheim, 9.–11. März 2011

vwh

Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Content Analysis in der Mathematik: Erschließung und Retrieval mathematischer Publikationen

Wolfram Sperber, Bernd Wegner

FIZ Karlsruhe – Zentralblatt MATH
Franklinstr. 11, 10587 Berlin

wolfram@zentralblatt-math.org
editor@zentralblatt-math.org

Zusammenfassung

Die traditionellen Informationsdienste in den Wissenschaften stehen angesichts der Publikationsflut und der Entwicklung der elektronischen Medien gerade auch in der Mathematik vor großen Herausforderungen. Es müssen neue Konzepte für eine erweiterte qualitativ hochwertige Erschließung entwickelt werden, die weitgehend automatisierbar sind.

In dem Beitrag werden die Konzepte zur Texterschließung in der Mathematik analysiert und Methoden entwickelt, die neue Möglichkeiten für das Retrieval eröffnen. Der Beitrag schließt mit einem Ausblick auf die Möglichkeiten zur Formel-Analyse.

Abstract

The classical scientific information services are confronted with big challenges: the increasing number of publications is calling for a new machine-based concept of content analysis and sophisticated methods for the retrieval. In the paper, a few new concepts for the content analysis and the retrieval of mathematical publications are presented. Moreover, the problem of formula analysis and retrieval is discussed.

1 Einleitung

Wissenschaftliche Publikationen bilden auch heute noch den Kern des Wissens in der Mathematik und haben eine Schlüsselrolle für das Auffinden und den Zugang zum mathematischen Wissen.

Mit dem Übergang in das industrielle Zeitalter und dem damit verbundenen Aufschwung in Forschung und Lehre hatte sich die Anzahl der wissenschaftlichen Publikationen sprunghaft erhöht. Für die Nutzer der Informationen, Wissenschaftler und Anwender, wurde es zunehmend aufwendiger und schwieriger, die zur Lösung eines Problems relevanten Publikationen zu identifizieren bzw. aufzufinden. In der Vergangenheit haben sich daher in verschiedenen wissenschaftlichen Disziplinen spezialisierte Informationsdienste, die Referatejournale, herausgebildet. Ziel der Referatejournale war (und ist es), den Lesern einen Überblick und eine Orientierungshilfe über die Entwicklungen in den Wissenschaften zur Verfügung zu stellen.

Die Forderung nach effizienten Werkzeugen für die Suche nach relevanten Informationen ist angesichts des ungebremsen Wachstums wissenschaftlicher Literatur und der Entwicklung der elektronischen Medien aktueller denn je. Insbesondere sind bessere Maschinen-basierte Methoden für die Erschließung der Literatur und die Einordnung der Ergebnisse in den wissenschaftlichen Kontext notwendig.

Universelle Suchmaschinen wie Google werden den Anforderungen aus den Wissenschaften nur zum Teil gerecht, da die Anforderungen und Interessen der Wissenschaften aus kommerzieller Sicht nur von untergeordnetem Interesse sind. In dem Beitrag werden für die Mathematik der Stand und die Perspektiven der inhaltlichen Erschließung mathematischer Literatur diskutiert.

2 Die Referatorgane und bibliografischen Datenbanken der Mathematik

Im 18. und 19. Jahrhundert stieg die Anzahl der wissenschaftlichen Publikationen immens an. Der erste eigenständige Referatedienst in der Mathematik, das Jahrbuch über die Fortschritte der Mathematik (JFM), wurde 1868 von Mathematikern in Berlin gegründet und umfasste 880 mathematische Publi-

kationen. Das JFM enthielt die bibliografischen Daten der Publikationen und häufig auch Besprechungen der Arbeiten, die von anderen Mathematikern auf freiwilliger Basis erstellt wurden.

1931 wurde in Deutschland ein weiterer Referatedienst für die Mathematik, das Zentralblatt für Mathematik (ZfM), gegründet. Es war insofern ein Gegenkonzept zum JFM, als dass die Aktualität absolute Priorität hatte und das Jahrgangsprinzip der JFM, also alle Arbeiten eines Jahres in einem Band zusammenzufassen und aufzubereiten, aufgegeben wurde. Mit Ende des 2. Weltkriegs wurde das JFM eingestellt. Bis in die 70-iger Jahre des 20. Jahrhunderts war sowohl die Produktion als auch das Produkt ZfM ausschließlich an das Papier gebunden. Mit dem Aufkommen der elektronischen Medien wurden zunächst die Produktion und dann auch das Produkt digitalisiert, es entstand die Datenbank ZBMATH, zunächst parallel zur gedruckten Form. 2010 wurde die gedruckte Form des ZfM eingestellt, der Nachweisdienst ZBMATH steht seitdem ausschließlich in elektronischer Form zur Verfügung. Heute ist die Datenbank ZBMATH der weltweit vollständigste und umfassendste Nachweisdienst für mathematische Literatur; für eine ausführliche Darstellung der mathematischen Referatedienste siehe die Artikel von (Wegner, 1998) und (Göbel & Sperber, 2010).

3 Content Analysis in den bibliografischen Datenbanken der Mathematik

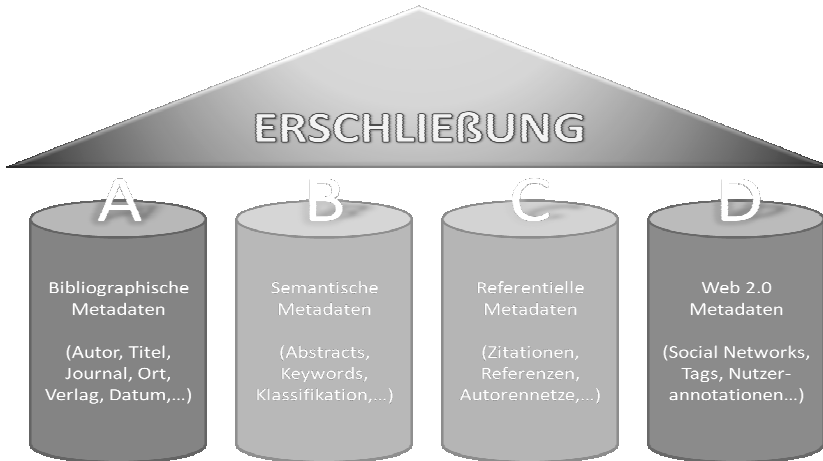
3.1 Qualitätsparameter der Referatedienste

Die Qualität und Attraktivität der Referatedienste macht sich an mehreren Kriterien fest:

- Vollständigkeit der Abdeckung des Gebietes, das durch den Dienst bearbeitet wird
- Umfang und Qualität der Erschließung des Materials
- Verknüpfung mit verwandten Informationen (Kontextbezug)
- Retrieval: Nutzerschnittstellen und Zugang zu den Diensten

3.2 Ein Modell für die Erschließung

Grob lässt sich die Erschließung wissenschaftlicher Publikationen den folgenden vier Kategorien zuordnen:



- Die *bibliografischen Metadaten* definieren das Minimum an Informationen, das für die Aufnahme in die Datenbank erforderlich ist, dazu gehören insbesondere Autor, Titel und Quelle.
- Die *referentiellen Metadaten* beinhalten die Verweise und Literaturreferenzen einer Publikation bzw. auf eine Publikation, aber auch Verknüpfungen mit anderen Autoren (Autorennetzwerke).
- Die *Web2.0-Metadaten* umfassen im Wesentlichen das Feedback der Community, etwa in Form von Kommentaren oder Diskussionsforen.
- Unter *semantischen Metadaten* werden alle Metadaten zusammengefasst, die Aussagen über den Inhalt oder dessen Einordnung in den wissenschaftlichen Kontext machen, insbesondere sind das der Abstrakt bzw. die Review einer Publikation, Keywords und Klassifikation. Diese Metadaten sind im Wesentlichen textbasiert. Abstrakt/Review, Keywords und Klassifikation haben eine eigenständige Bedeutung und sind komplementär zueinander.

Bibliografische, semantische, referentielle und Web2.0-Metadaten überlappen sich. Referentielle Metadaten etwa sind ein wichtiges Werkzeug für die Klassifizierung. Auch bibliografische Metadaten, etwa das Profil einer Zeitschrift, enthalten wichtige Anhaltspunkte über den Inhalt einer Arbeit.

3.3 Die Mathematical Subject Classification

In den 80er Jahren des 20. Jahrhunderts wurde von der American Mathematical Society (AMS) die AMS Subject Classification entwickelt und unter anderem für den Referatedienst Mathematical Reviews (MR) eingesetzt. Um dieses Schema als modernen Standard in die Mathematik einzuführen und weiter zu entwickeln, wurde Ende der 80er in Kooperation zwischen MR (heute Datenbank MathSciNet) und ZBMATH im Rahmen einer vertraglichen Vereinbarung eine gründliche Revision durchgeführt und diese fortan unter dem Namen Mathematical Subject Classification (MSC) weiterentwickelt. Die MSC ist ein hierarchisches dreistufiges System mit ca. 6.000 Klassen. Neben den hierarchischen Relationen zwischen Klassen existieren zwei weitere Arten von Verweisen zwischen den Klassen: „See also ...“ für Klassen ähnlichen Inhalts und „For .. see ...“ als Verweise auf Klassen, die spezielle Aspekte vertieft behandeln. Für weitere Informationen zur Klassifikation siehe (Göbel & Sperber, 2010).

Mit der Aktualisierung der MSC im Jahr 2010 wurde erstmals eine elektronische Master-Version eingeführt. Die Master-Version ist im TeX-Format, aus dem andere Formatierungen, etwa PDF, Word oder ein KWIC Index generiert werden. Die TeX-Version ist im Wesentlichen identisch mit der gedruckten Version. Semantische Aussagen, insbesondere die Relationen, sind nicht in Maschinen-verstehbarer Form dargestellt.

Die MSC weist zudem einige Schwächen im Design auf:

- *Definitionen der Klassen:* Die Definition der MSC-Klassen erfolgt ausschließlich über die Benennung der Klassen und deren Einordnung in das Klassifikationsschema. So umfasst etwa die Klasse „34Dxx Stability theory“ alle Publikationen, die sich mit der Stabilität der Objekte der Klasse „34-XX Ordinary differential equations“ befassen, also mit der Stabilität gewöhnlicher Differentialgleichungen.
- *Unübersichtlichkeit:* Die große Anzahl der Klassen der MSC und die teilweise starke inhaltliche Überlappung der Klassen führen dazu, dass die Arbeiten mehreren Klassen zugeordnet werden können. Andererseits ist die MSC – trotz der großen Anzahl von Klassen – für eine spezifische Suche oftmals nicht ausreichend (zu grob).
- *Ungleiche Wichtung der Klassen:* Die Klassen der MSC unterscheiden sich sowohl in Inhalt und Form als auch in der Granularität. Letzteres führt u.a. dazu, dass die Anzahl der Arbeiten, die einer Klasse zugeordnet sind, sehr unterschiedlich ist.

- *Lokales Design versus globales Design:* Die Weiterentwicklung der MSC erfolgt primär nach lokalen Gesichtspunkten, d.h. es finden die Erfordernisse einzelner Gebiete (MSC-Klassen der Top Ebene) Berücksichtigung. Prinzipien für ein einheitliches Design der MSC, etwa Konsistenz in der Strukturierung des Schemas, spielen eine eher untergeordnete Rolle. So werden z.B. Anwendungen in der MSC sehr unterschiedlich gehandhabt, teilweise werden die Anwendungsbereiche direkt benannt, meist sind sie aber unspezifisch.
- *Die Klassen sind verschiedenen Typs:* Die Klassen umfassen mathematische Objekte (etwa Gewöhnliche Differentialgleichungen), qualitative Aspekte (etwa Stabilität) oder Lösungsmethoden (etwa Finite Differenzenverfahren), etc.

3.4 Keywords und kontrolliertes Vokabular

Keywords sollen charakteristische Terme der bzw. über die Publikation enthalten, d.h. charakterisieren sowohl den Inhalt als auch ordnen die Publikation in den mathematischen Kontext ein. Bisher gibt es für die Mathematik noch kein kontrolliertes Vokabular.

Unter einem kontrollierten Vokabular der Mathematik wird im Folgenden die Menge der verwendeten Terme (Mehrwortphrasen) verstanden, die durch intellektuelle oder maschinelle Methoden aus dem vorhandenen mathematischen Wissen extrahiert wird und die für die Mathematik repräsentativ ist. Das kontrollierte Vokabular ist untrennbar mit der Entwicklung der Mathematik verbunden, durchläuft also einen stetigen Prozess der Veränderung und vergrößert sich ständig. Eine zuverlässige Abschätzung über die Größenordnung des verwendeten Vokabulars gibt es bisher nicht.

In einer Voruntersuchung wurden die Keywords der Datenbank ZBMATH untersucht. Die Analyse ergab einige überraschende Befunde, u.a. auch Hinweise auf die zu erwartende Größenordnung:

- Durchschnittlich sind jeder Publikation 3 Keywords zugeordnet.
- Häufig werden die Labels der MSC Klassen als Keywords verwendet, die Keywords fallen sogar häufig mit den Labels der MSC Klassen zusammen.
- Die Anzahl der verschiedenen Keywords für jede der 63 MSC Top-Klassen liegt deutlich über 1.000, d.h. ein kontrolliertes Vokabular für die Mathematik wird mehr als 500.000 Phrasen umfassen.

4 Ansätze für die semantische Erschließung in der Mathematik

4.1 Semantic-Web-Technologien

Semantic-Web-Technologien beschäftigen sich mit dem Problem, Informationen so darzustellen, dass deren Bedeutung auch von Maschinen erfasst werden kann. Informationen im Web lassen sich dann automatisch auswerten und verknüpfen, was neuartige Möglichkeiten für die Suche und den Zugang zu den Informationen eröffnet.

Mit dem Semantic Web stehen Methoden für eine erweiterte semantische Erschließung von Informationen zur Verfügung: (Resource Description Framework (RDF), 2004) und (RDF Vocabulary Description Schema Language 1.0: RDF Schema, 2004) als allgemeine Ansätze für die semantische Annotation, (Ontology Web Language (OWL), 2009) und (Simple Knowledge Organization System (SKOS), 2009) für die Definition von Ontologien, Klassifikationssystemen und Thesauri.

RDF und RDF Schema: RDF ist ein Graphenmodell, das es erlaubt, Aussagen der Form ‚Subjekt – Prädikat – Objekt‘ zu formulieren (etwa die Person A ist Autor der Publikation P) und diese zu verknüpfen. Mit RDF Schema wird das Vokabular für die RDF Darstellung der Informationen bereitgestellt.

OWL und SKOS: OWL und SKOS setzen auf RDF auf, benutzen also das Graphenmodell von RDF und das Vokabular von RDF Schema. Schon RDF Schema bietet mit dem Klassenkonzept die Möglichkeit, hierarchische Beziehungen abzubilden. OWL und SKOS verfügen darüber hinaus über ein spezielles Vokabular für Thesauri, Klassifikationssysteme und Taxonomien. So lassen sich etwa die Klassen der obersten Ebene eines Klassifikationsschemas auszeichnen oder die Relationen zwischen Klassen präzisieren.

Speziell für die Darstellung und Beschreibung mathematischer Inhalte wurden XML-Sprachen entwickelt, die die Möglichkeit bieten, mathematische Formeln und Symbole zu analysieren und suchbar zu machen. Darauf wird in Abschnitt 5 näher eingegangen.

4.2 MSC und Semantic-Web-Technologien

Die Transformation der MSC in das Semantic Web erfolgt schrittweise. In einem ersten Schritt wurde die MSC mittels des SKOS/RDF-Schema Vokabulars dargestellt. Hierzu gehören die Definition des MSC Schemas, der MSC Klassen sowie der Relationen zwischen den Klassen. Mit einer 1:1-Übersetzung der MSC von TeX nach SKOS ist es aber nicht getan. Um die MSC stärker für das Retrieval nutzbar zu machen, soll die MSC in einem zweiten Schritt überarbeitet und um zusätzliche semantische Aussagen über die Klassen der MSC erweitert werden. Insbesondere sind vorgesehen

- eine Typisierung der Objekte der Klassen, dafür wird gegenwärtig ein Schema entwickelt
- eine Präzisierung der Definitionen der Klassen über den Aufbau eines kontrollierten Vokabulars, siehe dazu den Abschnitt 4.3
- eine Präzisierung der Relationen zwischen den Klassen, etwa der Transitivität der hierarchischen Relationen
- die Überarbeitung der Klassenbezeichner, die Einführung alternativer Klassenbezeichner und die Zuweisung multilingualer Labels
- die Entwicklung von Konkordanzen, z.B. zur DDC und UDC, die für eine Interoperabilität mit Bibliothekssystemen relevant sind
- die Verknüpfung der verschiedenen MSC-Versionen, um die Entwicklung der MSC verfolgen zu können
- die Verlinkung mit anderen Informationsdiensten, etwa Wikipedia, ArXiv

4.3 Kontrolliertes Vokabular

Der Aufwand für den intellektuellen Aufbau eines kontrollierten Vokabulars für die Mathematik ist bei der zu erwartenden Größenordnung zu aufwendig. Stattdessen müssen maschinelle Lernverfahren eingesetzt werden, deren Resultate dann allerdings intellektuell ausgewertet und überprüft werden müssen.

Als Ausgangspunkt lassen sich das Vokabular der MSC sowie weitere vorhandene kontrollierte Vokabulare in der Mathematik nutzen, etwa die (Encyclopaedia of Mathematics (EoM), 2002). In einem zweiten Schritt sol-

len zusätzlich die in der Datenbank ZBMATH vorhandenen Keywords ausgewertet werden. Das führt dann zu einer Anreicherung des Startvokabulars um Keywords, die ebenfalls eine Klassifizierung gemäß MSC haben. Zudem ist durch die Häufigkeit ihres Auftretens eine Wichtung der Terme gegeben. In einem dritten Schritt schließlich soll das Startvokabular zur Extraktion von zusätzlichem Vokabular aus mathematischen Texten eingesetzt werden.

Erste Tests zur zusätzlichen Extraktion von Keywords aus Abstracts mathematischer Publikationen wurden zusammen mit W. Gödert, FH Köln für zwei MSC-Klassen (Gewöhnliche Differentialgleichungen und Graphentheorie) mit der Open Source Software Lingo durchgeführt. Neue Begriffe sind zumeist Mehrwortgruppen, die durch Kombinationen aus existierenden Begriffen entstehen. Das geschieht durch Phrasenbildung aus einem gegebenen Vokabular entsprechend vordefinierter Regeln (die aber flexibel angepasst werden können). Die Wortlisten der extrahierten Phrasen müssen anschließend intellektuell gesichtet werden. Die Tests haben zu etwa 30.000 relevanten Phrasen für jede der beiden MSC-Klassen geführt.

Von zentraler Bedeutung ist die Zuordnung der Terme des Vokabulars zur MSC. Es lassen sich damit – neben dem kontrollierten Vokabular für die gesamte Mathematik – spezielle Vokabulare für jede MSC Klasse aufbauen. Diese Klassen-spezifischen Vokabulare ermöglichen Aussagen über die Korrelation der Klassen. Zudem können diese Vokabulare für die automatische Klassifizierung eingesetzt werden. In einer Charakterisierung der MSC-Klassen durch ein kontrolliertes Vokabular sehen wir einen natürlichen Arbeitsschritt für die automatische Klassifizierung von Publikationen. Den MSC-Klassen werden dabei gewichtete Vektoren von Termen zugeordnet, die die Klassen inhaltlich definieren und als Maß für die Einordnung einer Publikation in eine MSC-Klasse genutzt werden. Übliche Verfahren der Textklassifizierung, siehe dazu etwa den Übersichtsartikel von (Sebastiani, 2002), also der Aufbau von Wortlisten aus Volltexten durch Elimination von Stoppwörtern, Stemming, n-grams, etc., liefern für die Mathematik unbefriedigende Ergebnisse. Ein kontrolliertes Vokabular (also eine Art „Positiv“-Termliste) ist ein anderer Ansatz zur Ermittlung der relevanten Phrasen für eine automatische Klassifikation. Mit der hier vorgeschlagenen Methode entsteht gleichzeitig ein neues Werkzeug für eine qualitativ bessere Keywordextraktion als auch die automatische Klassifizierung.

Das kontrollierte Vokabular bietet zudem die Möglichkeit, Ähnlichkeiten zwischen Publikationen unterhalb der MSC-Ebene zu identifizieren, also ein

Clustering der Publikationen vorzunehmen. Das ermöglicht neue Retrieval-funktionalitäten, etwa die Suche nach inhaltlich ähnlichen Dokumenten.

Keywords sind heute für das Retrieval wichtiger als Klassifikationssysteme. Das liegt u.a. an der fehlenden Kenntnis der Klassifikationssysteme bei vielen Nutzern, aber auch an den Nutzergewohnheiten, die sich durch die universellen Suchmaschinen wie Google ausgeprägt haben und ohne die (explizite) Nutzung von Klassifikationssystemen auskommen.

Ein kontrolliertes Vokabular kann zudem von den Autoren als Werkzeug für die Verschlagwortung seiner Publikationen genutzt werden, etwa indem die Autoren ihre Publikation über eine Schnittstelle eingeben und eine Vorschlagliste für Keywords erhalten.

5 Ein Ausblick in die Zukunft: Formelanalyse

Mathematik besteht bekanntlich nicht nur aus Text, sondern auch aus Formeln und Symbolen. Mathematische Symbole und Formeln komprimieren Sachverhalte, die sonst häufig nicht mehr in natürlicher Sprache dargestellt werden können. Symbole und Formeln enthalten in sehr verdichteter Form semantische Informationen. Mathematische Symbole und Formeln können im Abstrakt/Review, den Keywords und auch im Titel auftauchen.

Mit der Entwicklung der Rechentechnik ist Software zur Lösung mathematischer Aufgaben entwickelt worden, etwa Computeralgebrasysteme zur Lösung von Gleichungen. Diese Software ist häufig sehr speziell und muss miteinander verknüpft werden, um ein konkretes Problem zu lösen. Es müssen Methoden und Standards entwickelt werden, um Interoperabilität verschiedener Systeme zu erreichen. Im Rahmen von XML wurden Methoden; Standards und Markup-Sprachen entwickelt, etwa (MathML, 2010) oder (OpenMath), mit denen Symbole und Formeln eindeutig und Maschinen-verstehbar dargestellt werden können.

Mathematische Formeln spielten für das Retrieval in gedruckten Publikationen keine Rolle. Die Schwierigkeiten für das Retrieval von Symbolen und Formeln sind vielfältig. Die mathematische Formelsprache hat ähnliche Schwächen wie die natürliche Sprache, etwa die unterschiedliche Verwendung von Symbolen oder der fehlende semantische Bezug. Mit der Entwicklung spezieller XML Sprachen für die Mathematik wurden die Voraus-

setzungen geschaffen, um Methoden und Werkzeuge für die Erschließung von Formeln und deren Retrieval zu entwickeln. Erste Methoden und Konzepte befinden sich in der Diskussion.

Literaturverzeichnis

- FIZ Karlsruhe und American Mathematical Society (2010). Mathematics Subject Classification MSC. <http://www.msc2010.org> (Retrieved January 15, 2011)
- Göbel, S., Sperber, W. (2010). Bibliographische Information in der Mathematik – Werkzeug zur inhaltlichen Erschließung und für das Retrieval, Forum der Berliner Mathematischen Gesellschaft, Band 12, 70–99
- Hazewinkel, M. (2002). Encyclopaedia of Mathematics, Springer-Verlag: Berlin, Heidelberg, New York. <http://eom.springer.de/> (Retrieved January 15 2011)
- OpenMath Society (2009). Open Math. <http://www.openmath.org> (Retrieved January 15 2011)
- Sebastiani, F. (2002) Machine learning in automated text categorization, ACM Computing Surveys 34(1), 1–47
- W3C (2004). OWL Web Ontology Language Reference. <http://www.w3c.org/TR/owl-ref/> (Retrieved January 15, 2011)
- W3C (2004). RDF Vocabulary Description Language 1.0: RDF Schema. <http://www.w3c.org/TR/rdf-schema> (Retrieved January 15, 2011)
- W3C (2004). Resource Description Framework (RDF). <http://www.w3c.org/RDF/> (Retrieved January 15, 2011)
- W3C (2010). Mathematical Markup Language (MathML) Version 3.0. <http://www.w3c.org/TR/MathML3/> (Retrieved January 15 2011)
- W3C (2010). SKOS Simple Knowledge Organization System. <http://www.w3c.org/2004/02/skos/> (Retrieved January 15 2011)
- Wegner, B. (1998). Berlin als Zentrum des Wissenschaftlichen Referatewesens in Begehr, Heinrich: Mathematik in Berlin: Geschichte und Dokumentation, 1. Halbband; Shaker, 607-628