

Bachelorarbeit

Exploration und Visualisierung von wissenschaftlicher Literatur auf Grundlage externer Informationssysteme

Sebastian Holzki

16.09.2014



Gottfried Wilhelm Leibniz Universität Hannover
Fakultät für Elektrotechnik und Informatik
Institut für Verteilte Systeme – FG Wissensbasierte Systeme

Vorgelegt von

Sebastian Holzki – Matrikelnummer 2843720

Bearbeitungszeitraum

16.05.2014 – 16.09.2014

Gutachter

Prof. Dr. rer. nat. Robert Jäschke

Prof. Dr. techn. Wolfgang Nejdl

Betreuer

Prof. Dr. rer. nat. Robert Jäschke

Erklärung

Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen/Hilfsmittel verwendet habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe. Darüber hinaus versichere ich, dass diese Abschlussarbeit, in der vorliegenden oder in ähnlicher Form, noch keiner anderen Prüfungsbehörde vorgelegen hat.

Ort, Datum

Unterschrift

Inhaltsverzeichnis

Zusammenfassung	1
1. Einleitung	3
2. Grundlagen der Visualisierung	5
2.1. Die menschliche Wahrnehmung	5
2.2. Motivation und Theorie	8
2.2.1. Planung einer Datenvisualisierung	10
2.2.2. Grafische Exzellenz	13
2.2.3. Integrität von Visualisierungen	19
2.2.4. Explorative Datenanalyse durch Visualisierung	20
2.3. Vertreter der Visualisierung	21
2.3.1. Diagramme in der Koordinatenebenen	21
2.3.2. Zusätzliche Dimensionen in der Koordinatenebene	25
2.3.3. Die Landkarte als Metapher der Visualisierung	26
2.3.4. Netzwerk-Visualisierungen mit Graphen	27
2.3.5. Storyline Diagramme	29
2.3.6. Schlagwortwolken	32
2.4. Resümee und Einordnung	34
3. Praktische Ansätze der bibliografischen Organisation und Visualisierung	35
3.1. Ansätze zur Visualisierung	35
3.1.1. Science of Science Tool (Sci ² Tool)	36
3.1.2. Cytoscape	39
3.1.3. CiteSpace	42
3.1.4. MapEquation Alluvial Generator	47
3.1.5. Weitere Systeme zur Visualisierung	52
3.2. Ansätze zur Recherche und Organisation	65
3.2.1. Mendeley	65
3.2.2. Google Scholar	69
3.2.3. Microsoft Academic Search	70
3.2.4. CiteSeer	74
3.2.5. Harzing's Publish or Perish	76
3.3. Das fehlende Puzzlestück	77
4. Schnittstellen zu externen Informationssystemen	79
4.1. Zusammenfassung der geeigneten Informationssysteme	80
4.2. Die API von Microsoft Academic Search	82

5. Anwendungsfälle im Rahmen der Exploration und Visualisierung	86
5.1. Visualisierungen	87
5.1.1. Verteilung von Publikationen über die Zeit	87
5.1.2. Zitationsnetzwerk mit Zeitdimension	89
5.1.3. Community Analyse	91
5.1.4. Schlagwortwolke	93
5.2. Suchfunktion	93
5.3. Detailansicht	95
5.4. Sammlungen von Publikationen	96
5.5. Anwendungsfälle mit bedingter Umsetzbarkeit und Erweiterungen	98
5.5.1. Erweiterung der Suchfunktion	98
5.5.2. Erweiterung der Detailansicht	99
5.5.3. Erweiterung der Sammlungen von Publikationen	99
5.5.4. Übergreifende Erweiterungen	100
6. Die Implementierung von AcaVis	101
6.1. Programmarchitektur	101
6.2. Externe Bibliotheken	104
6.3. Java-Pakete und wichtige Verzeichnisse	106
6.4. Wartung und Weiterentwicklung	107
6.5. Quelltexte und Ressourcen von AcaVis	109
7. Fazit und Ausblick	110
Literaturverzeichnis	112
Danksagung	115
A. Weitere Screenshots der Benutzeroberflächen	118
A.1. Suchfunktion	118
A.2. Sammlungen von Publikationen	121

Zusammenfassung

Die Recherche von Literatur und Informationen ist ein elementarer Baustein wissenschaftlichen Arbeitens. Das klassische Vorgehen in diesem Bereich hat sich aufgrund der Verfügbarkeit von digitalen Informationssystemen in den letzten Jahren stark verändert. Bedeutender denn je ist es, wichtige von unwichtigen Informationen zu unterscheiden und ein organisiertes Arbeiten sicherzustellen. Das derzeit übliche Vorgehen ist dahingehend noch nicht perfekt. Es werden im Laufe dieser Arbeit viele bestehende Potenziale aufgezeigt. Besonders im Fokus steht dabei die Zuhilfenahme von Visualisierungen bei der Recherche. Eine Darstellung und Evaluation bestehender praktischer Ansätze in diesem Bereich wird vorgenommen. Die daraus gewonnenen Erkenntnisse dienen dann als Grundlage der prototypischen Entwicklung einer Webanwendung für die Exploration und Visualisierung von wissenschaftlicher Literatur.

The research of literature and information is an important part of the scientific workflow. The traditional way of performing these tasks has changed dramatically because digital information systems have become more and more popular during the past few years. Distinguishing between important and unimportant information as well as organizing these information in an efficient way are the most challenging tasks here. We will see that existing approaches can be improved in several ways. A detailed analysis will show the opportunities of visualizations for this purpose. Furthermore, the major practical approaches in this area will be presented and evaluated. The drawn picture leads to the development of a new approach, combining techniques of research and visualization. Finally, a web-based software prototype will be developed and documented using the proposed strategy.

1. Einleitung

Im Verlauf nahezu jeder Tätigkeit mit wissenschaftlichem Anspruch spielt das Beschaffen von Informationen eine wichtige Rolle. Noch vor einer inhaltlichen Sichtung von vermeintlich relevantem Material müssen die Themengebiete ergründet und zugehörige Publikationen gesammelt werden. Neben der klassischen quantitativen Sichtung von Material in einer Bibliothek stehen für diese Aufgabe auch vielerlei digitale Informationssysteme mit kaum erfassbaren, zunehmend wachsenden Datenbeständen zur Verfügung. Die effektive Verwendung dieser mächtigen und hilfreichen Werkzeuge gestaltet sich oft schwierig. Die Herausforderung besteht darin, Datenbestände auf relevante Teile einzugrenzen, Themen und Publikationen sinnvoll zu organisieren, sowie eventuelle Zusammenhänge zu verdeutlichen.

Zur Organisation und thematischen Aufarbeitung von Publikationen existieren bereits verschiedene Lösungsansätze. Diese erfreuen sich großer Beliebtheit in der wissenschaftlichen Gemeinde. Jedoch wird das Mittel der Visualisierung bei der Recherche wissenschaftlicher Literatur von vielen kaum in Betracht gezogen. Es gilt deshalb zunächst darzulegen, aus welchen Gründen Visualisierungen im Allgemeinen und in Bezug auf die bibliografische Organisation Vorteile mit sich bringen. Daher werden die Eigenschaften der menschlichen Wahrnehmung und Konzepte der Visualisierung, die diese Eigenschaften zu ihrem Vorteil nutzen, eingehender betrachtet. Weiterhin werden die hilfreichen Fälle für den Einsatz von Visualisierungen begründet. Dazu werden die bekanntesten Vertreter der Visualisierung kurz vorgestellt.

Die bekanntesten bestehenden Systeme zur Organisation und Aufarbeitung von Publikationen sollen in einem zweiten Schritt näher betrachtet werden. Die Systeme der bibliografischen Organisation werden bereits von einer breiten Öffentlichkeit genutzt. Daneben finden die Ansätze der unterstützenden Visualisierung von bibliografischen Informationen derzeit noch wenig Beachtung. Die zugehörigen Programme sind häufig komplex in ihrer Bedienung und leiden unter der schlechten Verfügbarkeit strukturierter und verwertbarer bibliografischer Informationen. Es lassen sich in diesem Zusammenhang zwei Lager identifizieren. Auf der einen Seite stehen die Systeme der Organisation und Recherche mit einer Datenbasis. Diese bereiten die vorhandenen Daten aber kaum mittels Visualisierungen auf. Auf der anderen Seite stehen Systeme der Visualisierung von bibliografischen Informationen, die aber ohne eine Datenbasis und Kenntnis der zugrunde liegenden Konzepte kaum nutzbar sind. Eine Auswahl der Systeme aus beiden Lagern wird bezüglich ihrer Vorteile und Schwächen untersucht. Dadurch entsteht ein übergreifendes Bild der gegenwärtigen Situation in Bezug auf die Recherche wissenschaftlicher Literatur und die bestehenden Möglichkeiten von Visualisierungen in diesem Bereich.

Auf Basis der gewonnenen Erkenntnisse wird dann ein eigener Lösungsansatz für die skizzierte Problemstellung erarbeitet. Dieser strebt die Möglichkeit einer Recherche wissenschaftlicher Literatur mit der Unterstützung von Visualisierungen an. Die aufgezeigte

Zweiteilung der bestehenden Systeme soll damit durchbrochen werden. Zunächst werden dafür die zweckmäßigen Anwendungsfälle und Visualisierungen entworfen. Dabei wird auf die Ergebnisse weiterer Arbeiten zu diesem Thema zurückgegriffen. Die dort vorhandenen Ansätze gelungener Visualisierung finden in dieser Arbeit Berücksichtigung. Die identifizierten Anwendungsfälle und Visualisierungen werden dann in einer prototypischen Webanwendung mit dem Namen *AcaVis* umgesetzt. Die besondere Eignung einer Webanwendung gegenüber Desktopanwendungen für den vorliegenden Fall wird ebenfalls dargelegt. Die Webanwendung soll bibliografische Daten aus bereits bestehenden Informationssystemen beziehen. Deshalb werden auch einige ausgewählte bibliografische Informationssysteme auf ihre Eignung hin untersucht. Das am besten geeignete System wird für die Implementierung des Prototypen ausgewählt.

Ergänzend zur Implementierung des Prototypen wird dessen Struktur abschließend dargestellt. Dies soll eine Fortentwicklung einzelner Teilbereiche beziehungsweise den Umbau für den Produktiveinsatz durch Dritte ermöglichen. Besonderer Wert wird bei der Darstellung auf die Programmarchitektur gelegt. Aber auch der Rückgriff auf externe Softwarebibliotheken wird beschrieben. Abschließend werden dann Einstiegspunkte für die Wartung und konkrete Weiterentwicklung der Webanwendung erläutert.

2. Grundlagen der Visualisierung

Ein wichtiges Ergebnis dieser Arbeit soll es sein, Daten zu visualisieren. Daten bilden die Basis von Information, Folgerungen bauen darauf auf und liefern die Erkenntnisse. Doch warum sollen gerade die Visualisierungen einen entscheidenden Vorteil beim Erkenntnisgewinn liefern? Um diese Frage zu beantworten, gebe ich zunächst einen Einblick in die menschliche Wahrnehmung allgemein und speziell bezüglich der visuellen Wahrnehmung. Wie sich herausstellen wird, sind unsere naturgegebenen körperlichen Eigenschaften ursächlich dafür, dass die Visualisierungen für uns so intuitiv und verständlich sind.

2.1. Die menschliche Wahrnehmung

Die hochentwickelten kognitiven Systeme sind eines der Merkmale, welche die Gattung Mensch besonders auszeichnen. Wahrgenommen werden vom Menschen die fünf Sinnesindrücke: Sehen, Hören, Fühlen, Riechen und Schmecken. Dabei sind die Sinnesorgane in ihrer Komplexität eher durchschnittlich, verglichen mit anderen Arten. Es ist die Verarbeitung der aktuellen Reize zusammen mit vergangenen Eindrücken, welche die menschliche Wahrnehmung erfolgreich macht. Hinzu kommt, dass der Mensch in der Lage ist, diese Fähigkeiten gezielt zu trainieren und sich so an verschiedene Umstände anzupassen. Auch befähigt die differenzierte Wahrnehmung seiner Umwelt den Menschen zur Kommunikation – ein entscheidender evolutionärer Vorteil. Was nützt mühsam erlangtes Wissen, wenn dieses gerade eine Generation überdauert?

Zur Kommunikation benutzt der Mensch im Allgemeinen fast ausschließlich die Sinnesindrücke des Sehens und des Hörens. So haben sich für diese Sinneseindrücke die umfangreichsten Zeichensysteme zur Kommunikation entwickelt. Kommunikation bezeichnet ganz allgemein den Austausch von Information zwischen mindestens zwei Stellen. Im gebräuchlichen Sinne bedeutet die Kommunikation zwischen zwei Menschen einen bidirektionalen Austausch von Information. Dabei können beide Stellen Informationen artikulieren und wahrnehmen. Als Kommunikation kann es aber auch verstanden werden, wenn eine Person Information von einer nicht-menschlichen Stelle erhält. Beispielsweise ist das Lesen eines Briefs oder das Betrachten eines Bildes ein Austausch von Information und somit Kommunikation. Es handelt sich jedoch in dem Moment um eine unidirektionale Kommunikation, denn dem Verfasser des Briefs oder des Bildes kann nicht unmittelbar eine antwortende Information gegeben werden.

Für den Bereich der Visualisierung ist diese Form von unidirektionaler Kommunikation vorherrschend. Missverständnisse in der unidirektionalen Kommunikation können im Allgemeinen nicht direkt durch Nachfragen aufgeklärt werden. Deshalb ist es wichtig, dass

die gesamte Visualisierung möglichst unmissverständlich ist und keine Fragen offen lässt. Im Detail ist die Umsetzung allerdings eine Herausforderung.

Obwohl die akustische und die visuelle Wahrnehmung in ähnlicher Art und Weise zur Kommunikation befähigen, gibt es zwischen ihnen einen bedeutenden Unterschied. Die Variation eines akustischen Signals wird vom menschlichen Gehör über die Zeit nur in einer Dimension erfasst (Bertin, 1983, S. 3). Ein visuelles Signal wird in insgesamt drei Dimensionen erfasst und die visuelle Wahrnehmung ist darüber hinaus in der Lage, Zusammenhänge zwischen den drei erfassten Dimensionen ohne Zeitverzögerung zu erkennen und Unterschiede zu differenzieren (Bertin, 1983, S. 3). Die Gegenüberstellung wird durch die Abbildung 2.1 verdeutlicht. Die visuelle Wahrnehmung stützt sich demnach auf eine deutlich komplexere Grundlage an Informationen als die auditive Wahrnehmung. Diese Eigenschaft verleiht ihr eine übergeordnete Bedeutung. Hinzu kommt, dass visuelle Informationen in Schrift und Bild dauerhaft festgehalten werden können. Diese Eigenschaft macht sich der Mensch bereits seit vielen Tausend Jahren zunutze (Bertin, 1983, S. 3f.). Demgegenüber ist die dauerhafte Speicherung von Audiosignalen erst seit gut Hundert Jahren möglich. Neben weiteren Aspekten ergibt sich daraus, dass die visuelle Wahrnehmung eher geeignet ist, Strukturen mit hoher Komplexität und einer großen Menge an kommunizierter Information zu erfassen.

System of perception		
		
sensory variables	1 variation of sound	1 variation of marks
	1 variation of time	2 dimensions of the plane
total	2 variables	3 variables
instantaneous perception	1 sound	Relationships among 3 variables

Abbildung 2.1.: Visuelle Wahrnehmung gegenüber auditiver Wahrnehmung
Quelle: Bertin (1983, S. 3, Abbildung 2)

Nicht jeder visuelle Eindruck wird gleichartig wahrgenommen. Teil des kognitiven Systems ist es auch, die Eindrücke zu verarbeiten. Dabei wird das wahrgenommene Bild mit bekannten Bildern aus dem Langzeitgedächtnis abgeglichen und es werden Muster verschiedener optischer Attribute wie Farbe und Form wiedererkannt. Dabei genügt es, wenn auch Bildausschnitte nur ähnlich sind. Diese partiellen Ähnlichkeiten werden bis zu einem bestimmten Grad wiedererkannt (Bertin, 1983, S. 2). Wo diese Grenze liegt, hängt in weiten Teilen vom Kenntnisstand und Training des Betrachters ab.

Besonders gut werden sehr homogene und inhomogene Bildausschnitte wahrgenommen, relativ zum visuellen Gesamteindruck. Dabei enthalten die besonders inhomogenen Bildausschnitte meist mehr Information als die homogenen. Dieser Aspekt wird durch Abbildung 2.2 verdeutlicht. Dem Betrachter erschließt sich sofort, dass es sich um die Silhouette einer Glühlampe handelt.

Betrachten wir aber die visuellen Attribute der Abbildung genauer. Dazu gehören beispielsweise Farbe, Kontrast und Struktur. Diese primitiven Attribute erschließen sich dem Betrachter ohne weitere Vorkenntnisse (Bertin, 1983, S. 2). Charakteristisch in der Abbildung ist es, dass die Silhouette durch einen starken Kontrast vom Grund abgehoben ist. Der Grund selbst kann grob als relativ homogene graue Fläche bezeichnet werden. Diese Eigenschaft wird durch die Verarbeitung des Bildes im Gehirn automatisch als unwichtig eingestuft, weil die Silhouette bereits einen größeren und offensichtlicheren Informationsgehalt geliefert hat. Die Folge dieser kognitiven Verarbeitung ist außerdem eine Unterteilung in Vordergrund und Hintergrund. Die Silhouette, mit größerem Informationsgehalt, liegt optisch im Vordergrund. Bei ganz genauer Betrachtung zeigt sich sogar, dass der Hintergrund gar keine homogene graue Fläche ist, sondern aus einem Muster verschiedener Grautöne besteht. Es handelt sich also um eine leicht inhomogene Fläche. Das zeigt: Die Ausprägung der Inhomogenität relativ zum Gesamteindruck ist ein wesentlicher Faktor für den Informationsgehalt der Abbildung.

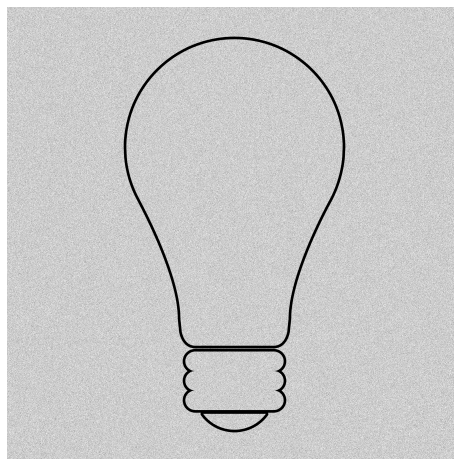


Abbildung 2.2.: Die Silhouette einer Glühlampe auf grauem Grund. Homogene und inhomogene Bildbereiche stechen heraus.
Quelle: eigenes Werk

Widmen wir uns jetzt der Semantik der grafischen Elemente in der Abbildung. Warum ist die Glühlampe eine Glühlampe und wird auch als solche erkannt? Dieses Erkenntnis beruht auf der Verknüpfung des wahrgenommenen Bildes mit Informationen aus dem Gedächtnis des Betrachters innerhalb des kognitiven Systems. Würde man umgekehrt eine Person bitten ein Bild einer Glühlampe zu malen, so würde dieses Bild der Abbildung 2.2 wahrscheinlich sehr ähnlich sehen. Notwendige Voraussetzung für die Verknüpfung mit bekannten Konzepten, wie dem der Glühlampe, ist aber deren Kenntnis durch den Betrachter. In Abbildung 2.3 ist die Silhouette eines menschlichen Kopfes zu sehen. Einem Betrachter mit besonderen Kenntnissen offenbart sich darüber hinaus, dass es sich um die Silhouette von Johann Wolfgang von Goethe handelt.

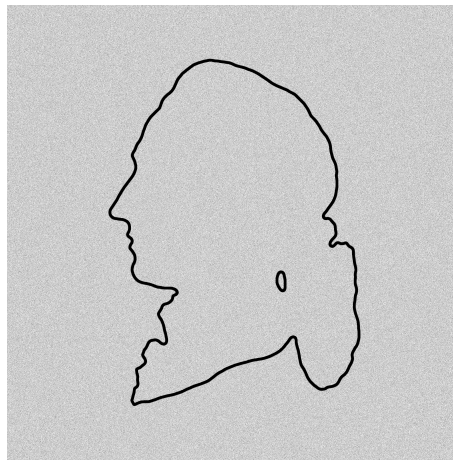


Abbildung 2.3.: Die Silhouette des Kopfes von Johann Wolfgang von Goethe. Durch Vorwissen unterscheidet sich, ob das Abbild durch den Betrachter als Kopf eines beliebigen Menschen oder als der Goethes erkannt wird.
Quelle: eigenes Werk

Die semantische Einordnung eines wahrgenommenen Bildes hängt also elementar vom Kenntnisstand des Betrachters ab (Bertin, 1983, S. 2). Ein Betrachter der die Abbildung betrachtet und die zugehörige Legende gelesen hat, wird diese Silhouette sicher als den Kopf von Johann Wolfgang von Goethe betiteln.

Das Studium des menschlichen kognitiven Systems fördert also Eigenschaften zutage, die wesentlich dafür sind, dass eine Kommunikation ohne Missverständnisse und Unklarheiten abläuft. Diese Eigenschaften lassen sich gezielt ausnutzen, um Vorgänge und Zusammenhänge auf eine Weise abzubilden, die mit Leichtigkeit wahrgenommen wird. Insbesondere soll die grafische Visualisierung von quantitativen Daten betrachtet werden, die im Gegensatz zur tabellarischen Darstellung deutliche Vorteile bietet.

2.2. Motivation und Theorie der grafischen Visualisierung

Grafische Visualisierungen haben sich als das Mittel der Wahl bewährt, komplexe Aussagen in einfachen Worten aus quantitativen Daten zu gewinnen. Der Charakter eines

Datensatzes wird nahezu ohne Verzögerung offensichtlich, sobald dieser angemessen visualisiert wird. Wie beschrieben, ist das die Folge der Funktionsweise der menschlichen Wahrnehmung. Ein minimales Beispiel veranschaulicht dies in der Abbildung 2.4. Die Datensätze I und III in tabellarischer Darstellung sind exakt, aber die Werte von besonderer Bedeutung bleiben hier verborgen. Während nach Studie der Tabelle zu I noch vermutet werden kann, dass es sich um streuende Werte handelt, bleibt verborgen, dass die Werte gut durch eine Gerade mit positiver Steigung gemittelt werden könnten. Aus der Tabelle zu III lässt sich kaum noch eine intuitive Aussage gewinnen. Der „ausreißende“ Wert lässt sich erahnen, der Charakter einer linearen Steigung ist vollkommen verborgen da die Tabelle nicht sortiert ist.

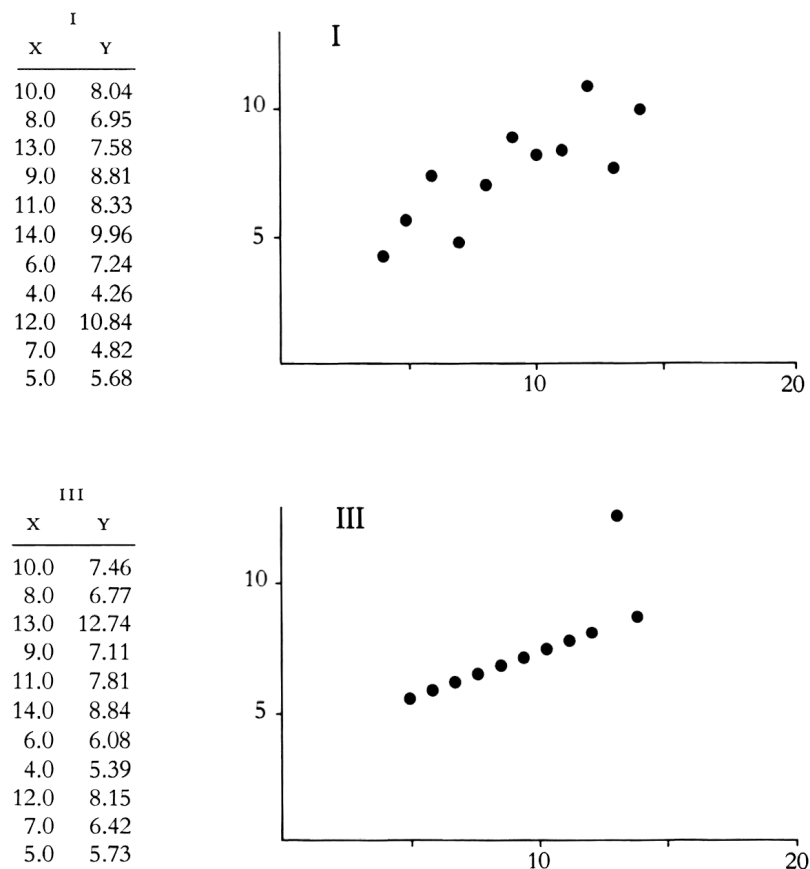


Abbildung 2.4.: Zwei fiktive Messreihen in tabellarischer Darstellung sowie als Diagramm. Die Messpunkte der Abszisse sind in I und III identisch.
Quelle: Tufte (1986, S. 13f.)

Hier sind lediglich einige der Eigenschaften der Datensätze beschrieben. Um die gesamte Charakteristik zu beschreiben, wäre weitaus mehr Text notwendig. Im Gegensatz dazu wird der Sachverhalt durch die Diagramme zu I und III kompakt und eindeutig dargestellt. Da die menschliche Wahrnehmung selektiv arbeitet, wird die Information sukzessiv

erfasst, je nach notwendigem Grad an Details. Wichtiger noch, es lassen sich die wesentlichen Aussagen der Diagramme – und damit auch der Daten – besonders kurz und knapp verbalisieren.

In diesem Beispiel lag die Visualisierung durch Diagramme sicherlich auf der Hand, doch nicht jeder Datensatz ist derart eindeutig in eine grafische Form zu bringen. Immens wichtig ist es daher zu erlernen, wie man die Werkzeuge der Visualisierung vorteilhaft und korrekt einsetzen kann. Die Vorgehensweise setzt sich aus den verschiedensten Disziplinen zusammen. Messmethoden, Statistik und künstlerische Fertigkeiten sowie das Verständnis der Zielgruppe sind nur einige dieser Disziplinen. Doch es ist keinesfalls notwendig, das Rad sprichwörtlich neu zu erfinden. Die bekannten Visualisierungen haben sich in den letzten Jahrhunderten entwickelt und bewährt (Tufté, 1986, Introduction). Die Vielzahl an verschiedenen Methoden der Visualisierung bietet ein reiches Portfolio, deren Elemente für den eigenen Zweck sinnvoll adaptiert und kombiniert werden können. Das Buch *Semiology of Graphics* (Bertin, 1983) bietet etwa auf über 400 Seiten Beispiele der einer Vielzahl von Visualisierungen – eine gute Möglichkeit um Anregungen zu sammeln. Mit dieser Menge an Möglichkeiten im Rücken liegt die Vermutung nahe: Für jeden Datensatz existiert eine grafische Darstellung, die diesen exakt visualisiert.

Die Vermutung mag stimmen, sie ist aber schier nicht zu bestätigen oder zu widerlegen. Grafische Visualisierungen dienen außerdem nicht dem Selbstzweck. Eine Visualisierung die kein Publikum hat, braucht nicht angefertigt werden. Das gebietet schon das Effizienzprinzip. So erübrigt sich schon die Frage nach einer passenden Darstellung für jeden existierenden Datensatz (Fry, 2008, S. 4, zweiter Absatz). Allgemein ist außerdem das bloße Sammeln und Archivieren von Daten wenig zielführend (Fry, 2008, S. 4, dritter Absatz). Wie im nächsten Abschnitt eingehender erläutert wird, sprechen rohe Daten ohne passende Aufbereitung in den seltensten Fällen für sich. Das Argument, die Daten würden für eine zukünftige Aufarbeitung zunächst erstmal gesammelt, wird durch folgende Effekte entscheidend geschwächt: Erstens ist der Datenbestand im Zweifel veraltet, sobald Ressourcen für die Aufarbeitung zur Verfügung stehen. Zweitens sind die gesammelten Daten höchstwahrscheinlich inkohärent mit einer im Nachhinein formulierten Fragestellung. Entweder sind viele irrelevante Informationen in den Daten enthalten, was die Handhabung erschwert. Oder, was noch fataler ist, es fehlen relevante Informationen in den Daten die nachträglich mühsam eingepflegt werden müssen. Ergo entspricht die Datenstruktur in solch einem Fall in vielerlei Hinsicht nicht den Anforderungen. Verhindern lassen sich diese Umstände mit einer durchdachten Planung. Es sei dennoch darauf hingewiesen, dass auch andere Ansätze ohne konkrete Zielführung existieren. Eine dieser Methoden wird zur Vollständigkeit im Abschnitt 2.2.4 kurz vorgestellt. Zunächst soll aber ein geplantes Vorgehen im Detail erörtert werden, da dieses eine Grundlage für das praktische Ergebnis dieser Arbeit bildet.

2.2.1. Planung einer Datenvisualisierung

Der zielführende Weg für eine gute Visualisierung und schlüssige Ergebnisse folgt einer vorherigen Planung (Fry, 2008, S. 5). Eine gute Planung basiert auf einer konkreten Fragestellung, die mit Hilfe von Daten und deren visueller Repräsentation beantwortet wird.

Die Fragestellung darf zunächst „High-Level“-Formulierungen enthalten. Das erleichtert die Interpretation der Ergebnisse, indem diese mit Hilfe der Visualisierungen quantitativ fundiert und in wenigen Worten beantwortet wird. Auch wird das Verständnis der Ergebnisse für die Zielgruppe leichter. Bestenfalls entfällt Fachterminologie völlig, wenn die Aussagen und Visualisierungen für sich sprechen. Umgangssprachlich geht es darum, im Vorhinein herauszufinden: „Welche Geschichte können die Daten erzählen?“ (Übersetzt nach Fry, 2008, S. 4, zweiter Absatz). Eine Motivation für die Fragestellung kann ein subjektiver Eindruck sein, ein gewisses Phänomen zu beobachten. Es geht dann darum, die Existenz des Phänomens objektiv zu verifizieren oder zu falsifizieren. Im Gegensatz zu einem formalen Beweis fußt diese Objektivität aber auf Daten, die in einem real existierenden System empirisch gesammelt werden (Fry, 2008, S. 2ff.). Eine Fragestellung muss aber keineswegs atomar und eng begrenzt sein. Vielmehr sollten bewusst Teilaspekte getrennt werden, um die Übersichtlichkeit zu wahren. Außerdem sollte eine Formulierung von der Umgangssprache über eine beschreibende Sprache in eine wissenschaftliche Fragestellung umformuliert werden. Manche Aspekte lassen sich schlichtweg nicht sinnvoll in einer allumfassenden Darstellung unterbringen. Aufteilung und getrennte Betrachtung (vgl. „Divide et Impera“¹) sorgen hier dafür, dass die Probleme leichter zu behandeln sind. Es ist einleuchtend, dass so für jedes Teilproblem nur die jeweils relevanten Daten betrachtet werden müssen. Umgekehrt wird also durch Nicht-Betrachtung irrelevanter Daten die Beantwortung der Fragestellung erleichtert.

Ein allgemeines Vorgehen von der Datensammlung, über die Verarbeitung, bis zur Visualisierung wird im Buch *Visualizing Data* (Fry, 2008) vorgeschlagen. Die dort genannten Stichpunkte können individuell ausgeführt werden. Es handelt sich nicht um den „heiligen Gral“ der Datenvisualisierung sondern lediglich um einen Leitfaden (Fry, 2008, S. 5f.). Nachdem im ersten Schritt die Fragestellung formuliert wurde, kann damit begonnen werden, die benötigten Informationen zu analysieren und eine treffende Datenstruktur zu entwerfen. Ein gründliches Vorgehen ist hier unerlässlich, weil die Informationen als Basis für die gesamte Argumentation dienen werden. Fehler in diesem frühen Stadium lassen sich später nur durch einen unnötigen Aufwand beheben. Es ist hilfreich, sich einem bewährten Datenmodell wie dem Entity-Relationship-Modell² zu bedienen. Insbesondere ist zu berücksichtigen, dass Informationen immer einem zeitlichen Wandel unterliegen. Daher muss geklärt werden: Genügt es, die Informationen als Momentaufnahme zu sammeln oder muss vielmehr ein mit der Zeit veränderlicher Strom an Informationen verarbeitet werden. Diese Überlegung hat deutlichen Einfluss auf die benötigte Datenstruktur.

Die Analyse der Datenquellen und das gezielte Sammeln der Informationen bleibt an dieser Stelle unbehandelt. Im Zusammenhang mit dieser Arbeit wird das praktische Vorgehen im Kapitel 4 beschrieben.

Vielfach enthalten die gesammelten Daten mehr Information, als sie im Rohformat preisgeben. Ein gutes Beispiel hierfür ist die statistische Aufarbeitung. Durch die Berechnung von Fehlermaßen oder Standardabweichungen lassen sich etwa besondere Eigenschaften

¹Divide et Impera (lateinisch für *teile und herrsche*) bezeichnet auch ein Vorgehen in der Problemlösung, bei dem ein Problem in kleinere Teilprobleme zerlegt wird, die für sich betrachtet leichter zu lösen sind.

²Das Entity-Relationship-Modell ist ein verbreitetes Schema zur Modellierung von Datenstrukturen. Die Hauptkomponenten sind Entitäten und ihre Attribute, sowie Beziehungen der Entitäten untereinander.

der gesammelten Daten in Zahlen ausdrücken. Diese allein sind schon teilweise aussagekräftig und können im Zusammenhang mit einer grafischen Visualisierung bestimmte Effekte verdeutlichen, die andernfalls nicht offensichtlich wären. Besonders bei der Analyse von Netzwerken ist auch die Berechnung von Metriken verbreitet, die Aussagen über die beteiligten Einheiten treffen. Diese Form von *Daten-Mining* hat unzählige Ausprägungen, nicht nur in der Statistik. Deshalb beschränke ich mich hier auf die Erwähnung der generellen Möglichkeit, durch Daten-Mining zusätzliche Informationen zu gewinnen.

Nachdem die Daten gesammelt, sowie durch Daten-Mining Informationen gewonnen wurden, müssen diese nun angemessen visualisiert werden. Die interessante Frage ist hier, welche Visualisierung den Daten und der Problemstellung Genüge tut. Eine Frage die pauschal nicht zu beantworten ist, denn sonst wäre dieser Abschnitt überflüssig. In den meisten Fällen offenbaren die Daten aber schon durch ihre Natur, welche Darstellungen auszuschließen sind.

Der Grundsatz, auf dem sämtliche Daten aufbauen, ist deren Variabilität. In der nachfolgenden Tabelle 2.2.1 ist ein exemplarischer fiktiver Datensatz aufgeführt. Dieser enthält drei unterschiedliche Komponenten mit ihrer individuellen Variabilität. Um alle diese Variablen gleichzeitig zu visualisieren würden also drei Dimensionen in der Darstellung benötigt.

Nachname	Position	Geburtsdatum
Maier	Abteilungsleiter	02.03.1975
Schmidt	Angestellter	20.08.1980
Seemann	Auszubildender	02.12.1989

Wichtiger ist aber die Ausprägung der Daten, die nach Bertin (1983) vier Formen hat: Selektivität, Assoziativität, Quantität und Ordnung. Mehrere dieser Ausprägungen können gleichzeitig auf eine Komponente der Daten zutreffen. Für die vorgestellte Tabelle ergibt sich daher eine klare Einordnung. Die Komponente *Nachname* ist selektiv aber nicht assoziativ, da die Namen unterschieden werden können, aber keiner gemeinsamen Gruppe angehören³. Die Komponente *Position* ist selektiv und assoziativ, weil beispielsweise alle Personen mit der Ausprägung *Angestellter* selektiv als Gruppe betrachtet werden können. Es ist möglich, dass auch eine Ausprägung von Ordnung vorliegt. Je nach Auffassung kann ein *Abteilungsleiter* als „größer“ als ein *Angestellter* betrachtet werden. Diese Wahrnehmung ist jedoch subjektiv und auf die semantische Klassifizierung der Begriffe zurückzuführen. Im Allgemeinen trifft die Einordnung deshalb nicht zu. Die dritte Komponente *Geburtsdatum* besitzt eine quantitative Ausprägung, da sie einen absoluten und relativen Wert besitzt. Aus diesem Grund ist sie auch geordnet. Weiterhin ist sie selektiv, da Personen mit unterschiedlichem Geburtsdatum unterschieden werden können⁴.

Sollen nun die Komponenten visualisiert werden, so liegt es auf der Hand, dass entsprechend der Zuordnung der Ausprägungen treffende grafische Attribute variiert werden

³Es wird davon ausgegangen, dass Personen mit dem selben Nachnamen nicht als Gruppe aufgefasst werden.

⁴Die Unterscheidbarkeit anhand des Geburtsdatums ist zwar im Allgemeinen nicht eindeutig, trotzdem ist die Ausprägung hier selektiv, entsprechend der vorliegenden Daten.

müssen. Auf dieser Grundlage können Visualisierungen, die das Profil nicht erfüllen, ausgeschlossen werden. Eine besondere Stellung unter den visuellen Attributen nimmt die Positionierung in der Ebene ein. Diese kann alle vier Ausprägungen für genau zwei Komponenten gleichzeitig erfüllen. So sollten die wichtigsten Bezugsdaten durch ihre Position in der Ebene dargestellt werden und im Anschluss die restlichen Komponenten durch passende Attribute wie Form, Farbe, Kontrast, Muster usw. ergänzt werden. Eine Überfrachtung ist jedoch zu vermeiden und im Zweifel die Visualisierung sinnvoll in ihren Komponenten zu unterteilen.

Die dargestellte Klassifikation lässt sich für sämtliche Formen der Visualisierung vornehmen. So kann ermittelt werden, ob ein Typ von Visualisierungen für die vorliegenden Daten theoretisch überhaupt geeignet ist. Das bedeutet jedoch noch nicht, dass diese Visualisierung auch die Aspekte treffend hervorhebt. Das Kriterium der „guten“ Eignung einer Visualisierung wird nach Tufte (1986) auch als *grafische Exzellenz* bezeichnet. Eine vorgenommene Visualisierung sollte vor der Veröffentlichung immer auf ihre Wirkung beziehungsweise ihre grafische Exzellenz geprüft und gegebenenfalls optimiert werden. Grafische Exzellenz ist durch verschiedene Faktoren messbar und es bestehen auch Methoden zur Verbesserung.

2.2.2. Grafische Exzellenz und Optimierung

Edward Tufte beschreibt den Begriff der grafischen Exzellenz in seinem Buch *The Visual Display of Quantitative Information* zunächst in aller Kürze und ohne einen konkreten Weg zu zeigen, wie grafische Exzellenz zu erreichen ist (Tufte, 1986, S. 51). Ein wesentlicher Aspekt dabei ist es, die Zielgruppe der Visualisierung zu kennen. Jeder Betrachter befindet sich auf einem individuellen Kenntnisstand von grafischen und didaktischen Konzepten. Eine Visualisierung soll griffige Aussagen möglichst unmissverständlich kommunizieren. Die Kommunikation muss zielgruppengerecht sein, das grenzt die Menge der in Betracht kommenden Visualisierungen ein. Tufte schreibt dazu außerdem sinngemäß übersetzt: „Eine Visualisierung ist grafisch exzellent, wenn sie dem Betrachter in kürzester Zeit die größte Anzahl an Ideen auf geringstem Platz mit möglichst wenigen grafischen Elementen liefert.“ (Tufte, 1986, S. 51). Siehe dazu auch die Abbildung 2.5. Das Schema zeigt den Betrachter und die Schlüsselaspekte, die eine gut kommunizierte Visualisierung ausmachen.

Es lässt sich nicht jede visuelle Wirkung exakt vorausplanen. Es kommt zu Situationen in denen sich herausstellt, dass die Visualisierung die gesetzten Ziele nicht umfänglich erfüllt. Dies kann zum Beispiel eine missverständliche Kommunikation der Hauptaspekte durch die Visualisierung sein. Bestenfalls wird dieser Umstand schon dem Autor selbst bewusst. Eine Garantie dafür besteht aber nicht, denn wie eingangs des Kapitels angeführt, hängt die Wahrnehmung immer von den persönlichen Attributen eines Betrachters ab. So ist es ratsam, die Darstellung absichtlich einer Prüfung durch unbeteiligte Probanden zu unterziehen, bevor diese veröffentlicht wird. Aus der Befragung der Probanden kann ein Soll-Ist-Vergleich angestellt werden, der gute und schlechte Eigenschaften der Darstellung charakterisiert. Insbesondere sollte darauf geachtet werden, dass eine möglichst große Zahl an Probanden mit verschiedenartig ausgeprägten persönlichen Attributen aus

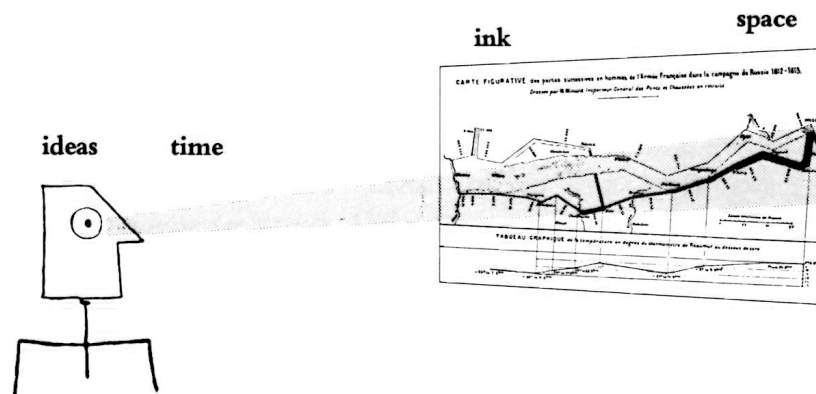


Abbildung 2.5.: Grafische Exzellenz nach Edward Tufte
Quelle: Tufte (1986, S. 51)

der Zielgruppe ausgewählt wird. Auf Basis der Ergebnisse kann dann eine Optimierung vorgenommen werden.

Besonders fatal ist eine schlechte Kommunikation zwischen der Visualisierung und dem Betrachter, wenn also der Betrachter die wichtigen Aussagen nicht erkennt oder diese sogar falsch interpretiert. Die Ursache dessen ist, dass viele Betrachter die „servierten“ Aussagen in großen Teilen hinnehmen, ohne eigenständig zu überprüfen. Im Grundsatz ist dieses Verhalten auch angebracht, sonst ginge der bedeutende Vorteil der Visualisierungen gegenüber anderen Kommunikationswegen verloren. Es ist also besonders wichtig, die Visualisierung bezüglich ihrer Aussagen vor der Veröffentlichung auf Verständlichkeit und Korrektheit zu prüfen. Daneben kann es auch weniger relevante Probleme geben, etwa die schlechte Wahl visueller Attribute wie Farben, Kontrast oder Symbolik. Dabei sollten die weniger relevanten Probleme keinesfalls unterschätzt werden. Priorität hat zwar die korrekte Kommunikation, doch in Summe kann es auch durch diese Probleme zu einer schlechten Kommunikation kommen.

Ein häufig vorkommender Fehler ist die Variation optischer Attribute, die für die jeweiligen Komponenten nicht geeignet sind. Die Eignung folgt, wie bereits angeführt, für jede Komponente aus ihren Ausprägungen. Verschiedene Farbtöne sollten niemals Daten visualisieren, die einer Ordnung folgen. Farbtöne sind von der natürlichen Wahrnehmung her ungeordnet. Im Allgemeinen sollte auf zu große Variation der Farbtöne verzichtet werden, denn die Anzahl möglicher Farbtöne ist begrenzt. Eine Verwendung von drei Farbtönen etwa lässt sich noch gut wahrnehmen, vor allem wenn diese sich an den drei Grundfarben rot, gelb und blau anlehnen. Um weitere Farben als die drei genannten zu erhalten, müssen die Grundfarben vermischt werden. Diese Mischfarben lassen sich schlechter unterscheiden, da sie durch die Verwendung der selben Grundfarben optisch ähnlicher wirken. Einen vergleichbar negativen Effekt hat die Verwendung zu vieler verschiedener Symbole. Je höher die Datendichte einer Visualisierung ist, desto kleiner muss die Repräsentation der einzelnen Datenpunkte ausfallen, um diese unterzubringen. Symbole müssen aber eine gewisse Mindestgröße besitzen, damit sie unterschieden werden können. Auch die Verwen-

dung zu vieler verschiedener Symbole beeinflusst die Wahrnehmung negativ. Die Symbole sind jeweils mit einer semantischen Bedeutung versehen, die sich von Visualisierung zu Visualisierung unterscheidet. Viele unterschiedliche Symbole erfordern ein häufiges Nachlesen in der Legende, was den Kommunikationsfluss behindert. Sehr gut wahrnehmbar sind hohe Datendichten jedoch mit anderen visuellen Attributen, wie dem Kontrast. Exemplarisch dafür ist in der Abbildung 2.6 eine Himmelskarte von Galaxien zu sehen. Neben der Anlehnung an eine geografische Karte ist darin als dritte Komponente die Anzahl der in einem Bildpunkt enthaltenen Galaxien abgebildet. Ein dunkler Grauton steht für eine große Anzahl, ein heller für eine kleine Anzahl von Galaxien. Insgesamt bildet die Karte circa 1,3 Millionen Galaxien ab (Bertin, 1983). Trotzdem lässt sich die Fülle an Informationen gut wahrnehmen, im Überblick wie im Detail.

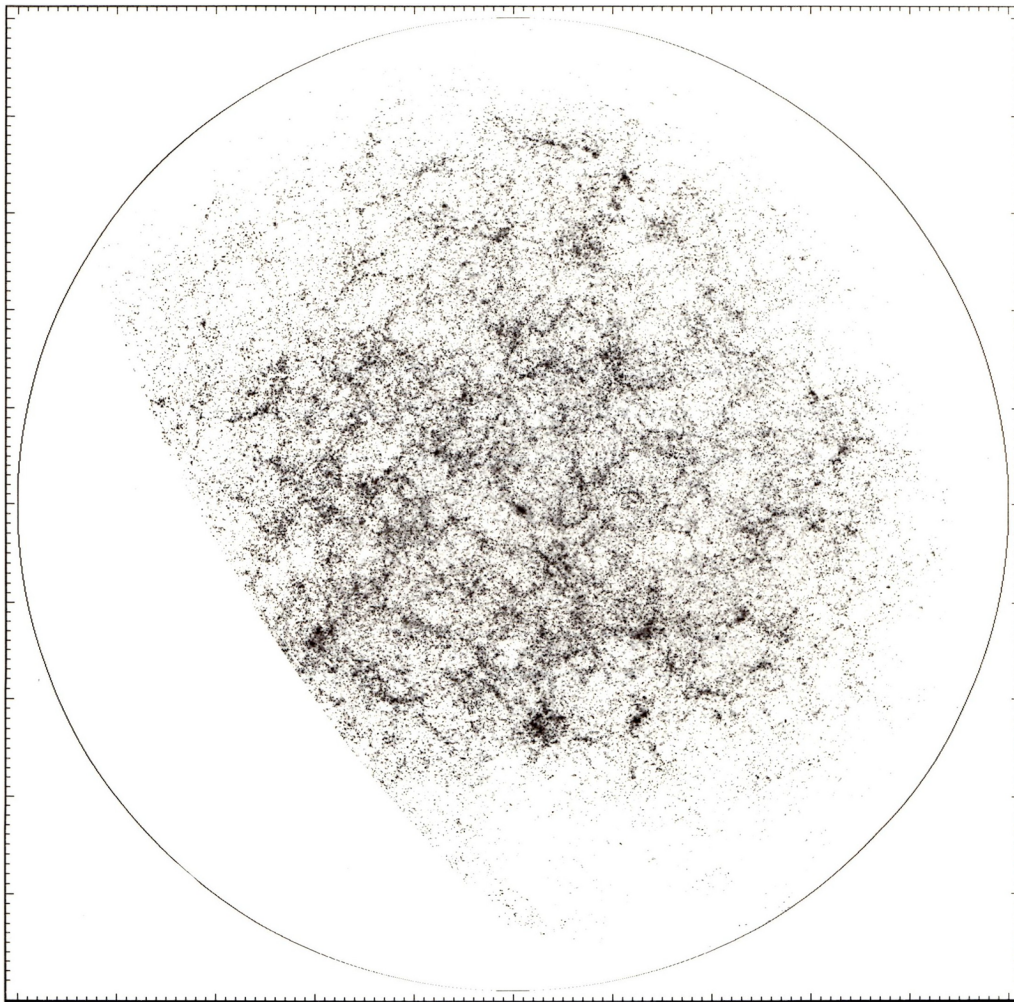


Abbildung 2.6.: Eine Visualisierung mit hoher Datendichte: 1,3 Millionen Galaxien als Himmelskarte
Quelle: Tufte (1986, S. 27)

Besonders intuitiv sind Visualisierungen, deren Komponenten in *natürlichen Bezugsgrößen* dargestellt werden. Das sind Raum, Zeit und Mengen. Jeweils in Größenordnungen, die im Alltag relevant und geläufig sind. Wann immer möglich sollten Visualisierungen auf diese natürlichen Bezugsgrößen zurückgreifen. Die Begründung liegt auf der Hand: Befinden sich Informationen zueinander im Raum, werden kognitiv automatisch Zusammenhänge wie Distanz, Nähe oder Häufung erfasst. Informationen im zeitlichen Wandel werden unterbewusst in eine Chronologie einsortiert. Eng mit dem zeitlichen Wandel verbunden ist auch das Erfassen von wiederkehrenden Ereignissen und Mustern in den Daten. Dimensionen in den natürlichen Bezugsgrößen erhöhen dagegen den kognitiven Aufwand bei der Erfassung vergleichsweise geringfügig. Es ist so möglich, vier oder fünf Dimensionen in einer Visualisierung darzustellen, ohne einen großen kognitiven Aufwand für den Betrachter zu erzeugen (Tufte, 1986, S. 40).

Ein Beispiel einer solchen Visualisierung mit natürlichen Bezugsgrößen ist in der Abbildung 2.7 zu sehen. Die Einzelheiten der zugrundeliegenden Statistik sind dem erklärenden Abschnitt in der Quelle, beziehungsweise dem Originalartikel⁵ zu entnehmen. Interessant ist die Visualisierung an dieser Stelle im Bezug auf die dargestellten Dimensionen. Die Basis bilden drei verschiedene Gruppen von Abgasen, jeweils gemessen in zwei Raumdimensionen, korrespondierend zur Landkarte in der Legende. Die Messung ist außerdem jeweils in vier Zeitabschnitte unterteilt, innerhalb derer die Messwerte summiert wurden. Insgesamt sind also fünf Dimensionen in einer einzigen Visualisierung untergebracht.

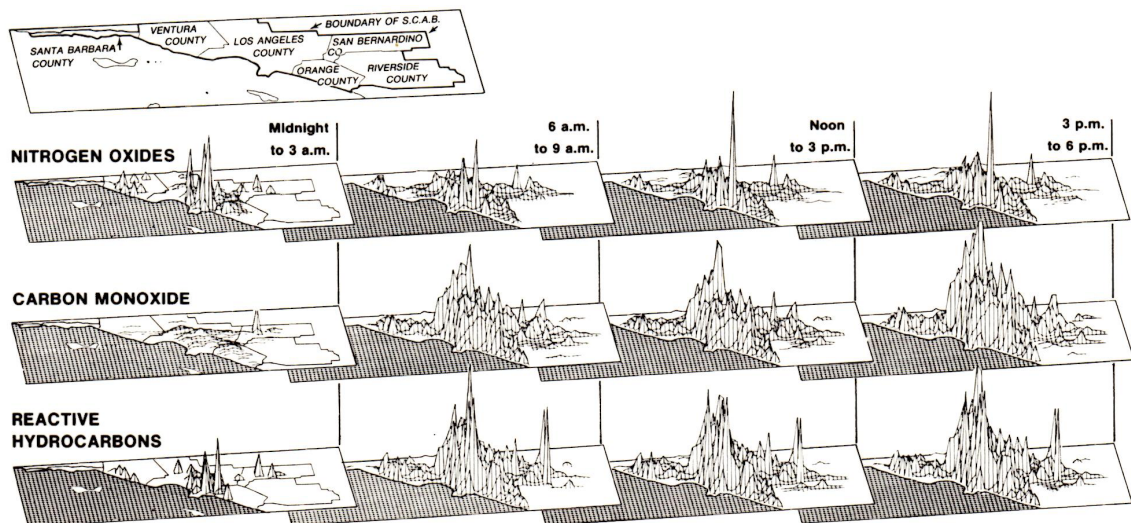


Abbildung 2.7.: Ausstoß von Stickoxiden, Kohlenstoffmonoxid und Kohlenwasserstoffen im Großraum Los Angeles zu verschiedenen Tageszeiten
Quelle: Tufte (1986, S. 42)

⁵Die Statistik und die zugehörige Visualisierung sind am 22. Juli 1979 in der Los Angeles Times erschienen (Tufte, 1986).

Das Konzept dieser Visualisierung ist nicht unbedingt geläufig, sodass zunächst ein gewisser kognitiver Aufwand des Betrachters notwendig ist, dieses zu verstehen. Im Anschluss ist es leicht möglich, einen Überblick über die Variation der Daten im Raum, entlang der Zeitdimension und der verschiedenen Abgase zu gewinnen. Die individuellen Eigenschaften der Visualisierung rücken in den Hintergrund, die Variation der Daten und Zusammenhänge einer höheren Ordnung, wie Ort und Zeit, werden deutlich. Mit etwas Hintergrundwissen können nach der Studie der Visualisierung leicht Aussagen über die zugrundeliegende Statistik abgeleitet werden. Auf das vorliegende Beispiel bezogen können das Aussagen über die Verursacher der verschiedenen Abgase sein, sofern eine Kausalität der Messwerte mit dem Ort und der Zeit als sinnvoll angenommen werden kann.

Edward Tufte stellt in seinem in seinem Buch *The Visual Display of Quantitative Information* auch eine Messmethode vor, die fehlende grafische Exzellenz hervorheben und den Erfolg von etwaigen Optimierungen quantitativ darstellen kann. Diese liefert als Resultat quantitativ die sogenannte *Data-Ink-Ratio* (Tufte, 1986, S. 93). Diese benennt einen realen Wert zwischen 0 und 1, der die dargestellten Daten und das Beiwerk der Visualisierung ins Verhältnis setzt.

$$\text{Data-Ink-Ratio} = \frac{\text{Data-Ink}}{\text{Total Ink}}$$

Die Begriffe sind historisch begründet und beziehen sich auf die Tinte (engl. Ink) die auf einen Träger aufgebracht werden muss, um die Visualisierung abzubilden. Das Konzept lässt sich jedoch vollständig auf digitale Visualisierungen übertragen, da die *Data-Ink-Ratio* dimensionslos ist. Die *Data-Ink* bezeichnet den Teil der „Tinte“ der zur Visualisierung der einzelnen Komponenten notwendig ist. Die *Total-Ink* bezeichnet die „Tinte“ zur Darstellung der gesamten Visualisierung. Daraus folgt unmittelbar: Die *Data-Ink-Ratio* ist 1, wenn ausschließlich Daten abgebildet sind und 0, wenn keine Daten abgebildet sind.

Viel grafisches Beiwerk, das keine Daten als wesentliche Information abbildet, reduziert die Aussagekraft der Visualisierung. Das Ziel ist es, eine möglichst hohe *Data-Ink-Ratio* zu erreichen, wobei ein Wert von 1 selten erreicht wird. Eine hohe *Data-Ink-Ratio* wirkt sich positiv auf die grafische Exzellenz aus, sofern diese begründet ist (Tufte, 1986, S. 93ff.). Um das Ziel zu erreichen, können zwei Maßnahmen ergriffen werden: Erstens das Hinzufügen von weiteren Daten, also eine Erhöhung der *Data-Ink* und zweitens das Entfernen von *Non-Data-Ink*, also dem Entfernen von grafischem Beiwerk (Tufte, 1986, S. 96). Die *Data-Ink* zu erhöhen ist nur bis zu einem gewissen Grad möglich, da die Anzahl der Daten begrenzt ist und die visuelle Erscheinung bei einer zu hohen Datendichte leidet. Viel häufiger kann das Optimierungspotenzial durch die zweite Maßnahme ausgenutzt werden. Hier ist jedoch darauf zu achten, dass es auch *Non-Data-Ink* gibt, die für das Verständnis der Daten zwingend notwendig ist. Diese darf auf keinen Fall entfernt werden. Edward Tufte schreibt dazu wortwörtlich: „Erase non-data-ink, within reason“ (Tufte, 1986, S. 96).

Ein anschauliches Beispiel mit variierender *Data-Ink-Ratio* innerhalb der gleichen Visualisierung ist in der Abbildung 2.8 zu sehen. Dargestellt sind die erwartete und eingetretene Wahlbeteiligung zur Präsidentschaftswahl 1960 in den Vereinigten Staaten pro Stadt, für insgesamt 104 Städte. Dabei übertrafen einige Städte die erwartete Wahlbeteiligung und

2. Grundlagen der Visualisierung

andere verfehlten sie. Punkte die exakt auf der eingezeichneten 45 Grad Achse liegen, repräsentieren demnach Städte, in denen die vorhergesagte Wahlbeteiligung exakt eingetreten ist. Mit Wahlbeteiligung ist in diesem Fall bereits die Registrierung zur Teilnahme an der Wahl gemeint. Die Visualisierungen wurden chronologisch (von links nach rechts) optimiert und veröffentlicht. Das linke Diagramm enthält sämtliche Datenpunkte, besitzt also maximale Data-Ink. Trotzdem fällt die Data-Ink-Ratio sehr niedrig aus, da die Datenpunkte zwischen den Hilfslinien komplett untergehen. Im mittleren Diagramm wurde ein Großteil der Non-Data-Ink entfernt, jedoch auch die Datenpunkte. Die Data-Ink-Ratio liegt deshalb bei 0. Im rechten Diagramm wurden die Datenpunkte hinzugefügt und eine weitere Reduzierung der Non-Data-Ink vorgenommen. Die Data-Ink-Ratio liegt deshalb bei etwa 0,7 (Tufte, 1986, S. 95).

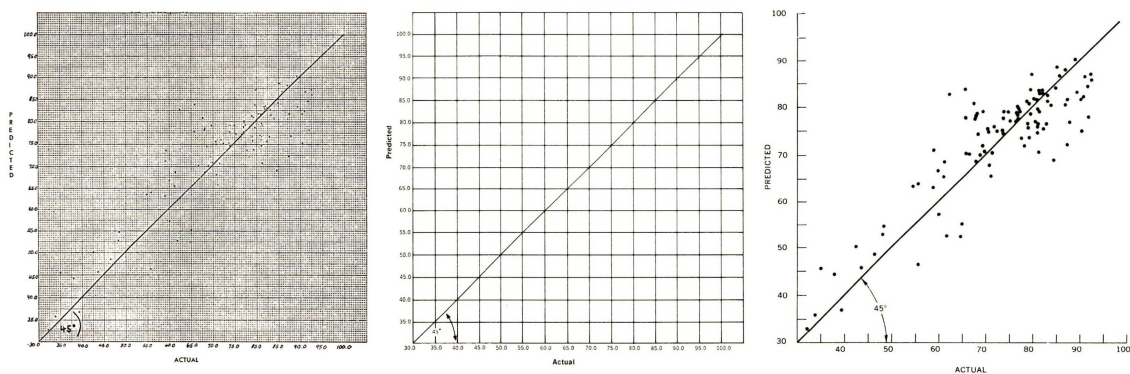


Abbildung 2.8.: Der selbe Datensatz in Diagrammen mit verschiedener Data-Ink-Ratio
Quelle: Tufte (1986, S. 94f.)

Eine weitere Reduktion der Non-Data-Ink im rechten Diagramm ist kaum möglich. Sie würde zwar die Data-Ink-Ratio steigern, die visuelle Wirkung aber verschlechtern, denn die Skalen und Beschriftungen sind für das Verständnis zwingend notwendig. Das zeigt also: die Messmethode der Data-Ink-Ratio trifft nur bedingt Aussagen über die echte grafische Exzellenz, kann jedoch ein Anhaltspunkt für die Optimierung von Visualisierung sein. Die Interpretation des quantitativen Wertes der Data-Ink-Ratio muss deshalb immer individuell vorgenommen werden.

Besonders fragwürdig im Hinblick auf die Data-Ink-Ratio sind auch Balken- und Tortendiagramme, die besonders häufig zur Darstellung statistischer Sachverhalte eingesetzt werden. Vielfach wird ein kleiner Datensatz durch riesige Flächen visualisiert. Das ist für die Lesbarkeit teilweise gerechtfertigt, dennoch handelt es sich um *redundante Data-Ink* (Tufte, 1986, S. 96, zweiter Abschnitt). Eine Optimierung der visuellen Attribute bewegt sich deshalb auf einem schmalen Grad zwischen Lesbarkeit, Ästhetik und Effizienz der Kommunikation. Ein Beispiel dazu ist der Abbildung 2.9 zu entnehmen. Edward Tufte stellt diese mit weiteren Zwischenschritten vor, die dargestellten Schritte I bis III sind dabei diejenigen mit den drastischsten Optimierungen. Das ursprüngliche Diagramm I beinhaltet ein typisches Balkendiagramm nicht näher benannter Daten, wie es zum Beispiel auch aus Tabellenkalkulationsprogrammen bekannt ist. Die Menge an grafischem

Beiwerk ist anhand der Data-Ink-Ratio zu identifizieren und wird deshalb in den Optimierungen von Diagramm I zum Diagramm II reduziert. Einige Elemente mit zusätzlicher Aussagekraft kommen dort auch hinzu. Schließlich wird im Diagramm III wieder ein Teil der Anpassungen weggelassen, da dies die Effizienz der Kommunikation zwischen Visualisierung und Betrachter nicht negativ beeinflusst. Die Einteilung in Schritten von je fünf Prozentpunkten wurde hier sogar durch eine reine Reduktion der redundanten Data-Ink erreicht.

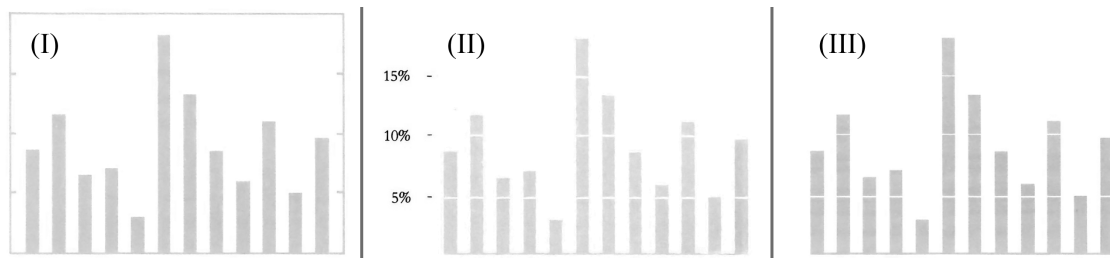


Abbildung 2.9.: Optimierung der Data-Ink-Ratio eines Balkendiagramms
Quelle: Tufte (1986, S. 126ff.)

Trotzdem ist noch ein Anteil an redundanter Data-Ink enthalten. Ob diese Balken noch zu schmalen Strichen reduziert werden könnten, hängt wieder von den oben genannten Faktoren ab. Anhand des Beispiels lässt sich also erkennen: Es gibt Typen von Diagrammen, die in der Lage sind, viele unterschiedliche Datentypen zu visualisieren. Doch das bedeutet nicht, dass die Visualisierung dadurch exzellent ist. Die Visualisierungen sollten immer entsprechend der individuellen Problemstellung optimiert werden, um diesen grafische Exzellenz zu verleihen. Der Schritt der Optimierung wird jedoch allzu häufig unterlassen.

Diese Zusammenstellung bietet nur einen kleinen Einblick in die Potenziale der Optimierung von Visualisierungen. Eine umfängliche Betrachtung würde im Rahmen dieser Arbeit zu weit führen. Deshalb fahre ich im folgenden Abschnitt fort mit der Integrität von Visualisierungen, welche im weitesten Sinne auch zur Exzellenz beiträgt.

2.2.3. Integrität von Visualisierungen

Fehlende Integrität wird häufig als negative Eigenschaft von Visualisierungen wahrgenommen. Dieser Eindruck wird dadurch gestärkt, dass viele Visualisierungen publiziert werden, die durch unsachgemäßes Vorgehen oder sogar mit Vorsatz Sachverhalte suggerieren, die trotz ihrer Plausibilität objektiv nicht der Wahrheit entsprechen. Eine Visualisierung dient dem Zweck, große Datenmengen und komplizierte Zusammenhänge zu offenbaren. Der Betrachter ist auf die Integrität der Visualisierung angewiesen, sofern er die Ergebnisse nicht unabhängig verifizieren möchte. Einer vorsätzlich falschen Darstellung ist nur durch unabhängige Überprüfung der Ergebnisse zu begegnen, ein Vorgehen dass schon immer die Praxis wissenschaftlicher Arbeit war. Besonders ärgerlich für den Autor einer misslungenen Visualisierung ist es, wenn diese nicht mit Vorsatz, sondern aus Unwissenheit entstanden

ist. Neben den Ansätzen zur grafischen Exzellenz, wie sie im vorigen Abschnitt genannt wurden, lassen sich ganz allgemein gewisse Regeln anwenden, um Fehler zu vermeiden.

Erstens sollten visualisierte Größen immer direkt proportional zu den realen Größen sein. Zweitens sollten niemals die optischen Eigenschaften zwischen Visualisierungen variieren, die zusammen gehören. Drittens sollten Daten bestenfalls innerhalb der Visualisierung anstelle eines beschreibenden Textes beschriftet werden. Viertens sollten die Daten immer im Vordergrund stehen, grafisches Beiwerk ohne Dateninhalt sollte soweit wie möglich vermieden werden. Diese vier Prinzipien wurden im Kern von Edward Tufte publiziert (Tufte, 1986, S. 77).

Auf eine Darstellung derartiger Negativbeispiele wird in dieser Arbeit bewusst verzichtet. Dieser Abschnitt dient lediglich der Sensibilisierung für das Thema mit Hinweis auf die wissenschaftliche Praxis. Mit welchen Mitteln mangelhafte Visualisierungen offenbart und die Aussagen überprüft werden können, wird in diversen Publikationen beschrieben. An dieser Stelle sei auf das Kapitel „Chartjunk“ im hier zitierten Buch von Tufte (1986) verwiesen.

2.2.4. Explorative Datenanalyse durch Visualisierung

Im Kontrast zum vorgestellten Verfahren der geplanten Datenvisualisierung existiert auch ein Konzept, welches ohne konkrete Fragestellung Aussagen generiert, die sogenannte *explorative Datenanalyse*. Ziel ist hierbei, nicht vorformulierte Aussagen über die Daten aus den Visualisierungen zu gewinnen, daher die Bezeichnung als „explorativ“. Das Verfahren soll hier nur kurz angerissen werden, da der Anspruch dieser Arbeit eine begründete und geplante Visualisierung auf Grundlage konkreter Fragestellungen ist. Siehe dazu auch Kapitel 4 und Kapitel 5. Die explorative Datenanalyse wurde von John Turkey um 1977 geprägt (Chatfield, 1986). Turkey verfolgte das Ziel, Datensammlungen weitgehend ohne konkrete Fragestellung anhand von bewährten Verfahren zu visualisieren (Fry, 2008). Die Grundlage, auf der Turkey die Vorgehensweise rechtfertigt ist, dass die interessanten Eigenschaften eines Datensatzes automatisch offensichtlich würden, wenn dieser visualisiert wird (Chatfield, 1986). Auch dienen Vermutungen, die ähnlich zu bereits untersuchten Phänomenen scheinen, als Anhaltspunkt für die Vorgehensweise (Chatfield, 1986). Die Methodik besitzt ihre Vorteile insbesondere in der statistischen Aufarbeitung von Daten, weil sich Verfahren und Kenngrößen etabliert haben, welche quantitative Anhaltspunkte für den Charakter liefern (Fry, 2008, S. 4, vierter und sechster Absatz). Fraglich ist dennoch, welche Visualisierung jeweils das Mittel der Wahl ist, auch wenn es Erfahrungswerte und grobe Anhaltspunkte gibt. Insbesondere stellt sich diese Frage bei Visualisierungen, die nicht durch die Statistik motiviert sind. Das Problem existiert analog zur geplanten Datenvisualisierung. In diesem Kontext wurde bereits vorgestellt, wie mit dem Charakter der Daten und den formulierten Vermutungen auf infrage kommende Visualisierungen geschlossen werden kann.

Wer den Weg der Visualisierung wählt, muss sich seines Ziels im Klaren sein. Während die geplante Visualisierung die Existenz eines Phänomens untermauern soll, möchte die explorative Datenanalyse Phänomene ausfindig machen. Bei der eigenen Umsetzung ist es daher wichtig, sehr gut darüber nachzudenken, um welche Form von Problemstellung

es sich handelt. Eine Fragestellung in den explorativen Kontext zu rücken, nur weil unter dem Mantel des Modebegriffs *Big Data* ein großes Daten-Sammelsurium archiviert wurde, ist wenig zielführend, denn Daten sollten immer zielorientiert erhoben werden.

2.3. Vertreter der Visualisierung

Jede Visualisierung besteht aus mindestens einem der folgenden drei Grundelemente, häufig aus Mehreren:

- **grafische Elemente** z.B. Achsen, Raster, Datensymbole
- **Texte** z.B. Überschrift, Beschriftungen, Randparameter, Zusammenfassung
- **Tabellen** z.B. Datentabelle und Legende

Den größten Anteil der Visualisierung sollten zwar Grafik und Symbolik ausmachen, die anderen Elemente sind deshalb aber nicht weniger wichtig. Mehr noch sind sie unerlässlich für die unmissverständliche Aussage der Visualisierung für den Betrachter (Tufté, 1986, S. 77). Jede Visualisierung muss in einen Kontext eingebunden sein (Tufté, 1986, S. 77). Zum Kontext gehören alle Informationen, die nicht aus den erhobenen Daten folgen. Unverzichtbar sind ein prägnanter Titel für die Visualisierung, das Hervorheben von Datenpunkten mit hohem Aussagewert, sowie die Maßeinheiten der dargestellten Informationen. Weniger offensichtlich, aber unbedingt zu nennen, sind die *Invariablen* unter denen die gezeigten Daten erhoben wurden. Das sind zum Beispiel die Randparameter, die im Laufe einer Messreihe konstant bleiben oder eine Aussage, zu welcher bestimmten Zeit eine Statistik wiederkehrend erhoben wird. Werden diese Grundregeln missachtet, so können die Ergebnisse nicht eindeutig nachvollzogen werden. Es ist dabei immer Rücksicht auf die Zielgruppe mit der größten Unkenntnis, bezogen auf den Sachverhalt, zu nehmen.

Die aufgeführten Informationen sollten immer explizit genannt werden, selbst wenn sie dem Autor redundant oder überflüssig erscheinen. Nur in absoluten Ausnahmefällen und mit guter Begründung können diese wirklich weggelassen werden. Einen solchen Fall stellt die Verwendung einer guten Metapher zur Visualisierung dar. Dabei wird beispielsweise das Konzept einer Landkarte dazu verwendet, Daten mit einem örtlichen Bezug zu dieser Landkarte zu visualisieren. Wichtig bei der Verwendung ist es, sicherzugehen, dass den Betrachtern jeder Zielgruppe die Metapher bekannt ist.

Ergänzend zu diesen theoretischen Eigenschaften und den beispielhaften Visualisierungen aus den vorangegangenen Abschnitten des Kapitels werden nun noch verschiedene Vertreter der Visualisierung vorgestellt. Diese sollen erste Anregungen für eine Umsetzung der Visualisierungen im praktischen Teil dieser Arbeit geben.

2.3.1. Diagramme in der Koordinatenebenen

Die bekanntesten Vertreter der Visualisierung sind Diagramme in der Ebene, die ein Koordinatensystem als Bezugssystem verwenden. Diese lassen sich in drei grundlegende Typen unterteilen die bestimmen, wie die Daten im Diagramm dargestellt sind. Daten lassen sich

in der Ebene als Punkte, Linien und Flächen darstellen (Bertin, 1983, S. 44). Zunächst betrachten wir die einfachste Darstellungsform, den Datenpunkt. Er hat in der einfachsten Form keine Ausprägung oder Dimension, auch wenn er durch ein beliebiges Symbol dargestellt werden kann. Neben der eigentlichen Positionierung in der Ebene ist für die Aussage von Punktdaten entscheidend, in welchem Verhältnis sich diese zueinander befinden. Insbesondere im Zusammenhang mit der natürlichen Bezugsgröße Zeit können so leicht Aussagen und Trends abgeleitet werden. Diagramme mit der Bezugsgröße Zeit werden speziell auch *Histogramme* genannt. Eine solche Darstellung von Punktdaten mit zeitlicher Bezugsgröße ist in der Abbildung 2.10 dargestellt. Die diskreten Werte wurden jeweils zum Ende des vierten Quartals eines Jahres aufsummiert.

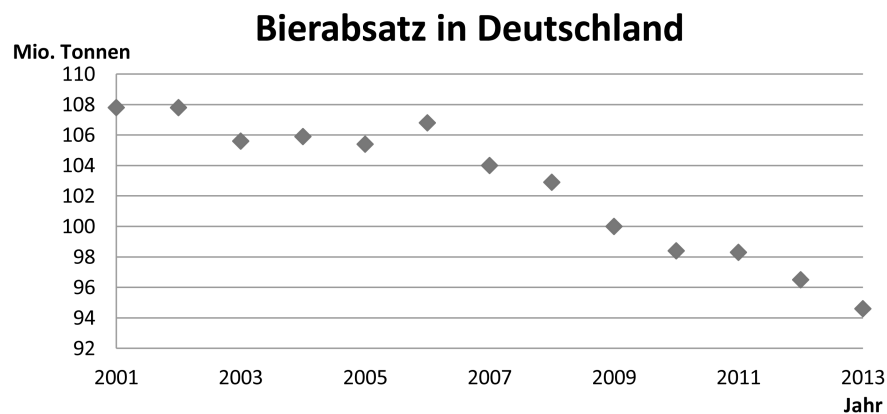


Abbildung 2.10.: Bierabsatz in Deutschland zwischen 2001 und 2013
Datenquelle: DESTATIS – ohne Autor (2014b, Tabelle)
Visualisierung: eigenes Werk

Beide Messgrößen, die Zeit in Jahren und die Menge in Tonnen, sind natürliche Bezugsgrößen und intuitiv verständlich. Sie bewegen sich außerdem in Größenordnungen, die im Alltag geläufig sind. Die Aussage des Diagramms und der Daten wird jedoch noch nicht deutlich genug. Als Maß für diese Aussage kann der subjektive Eindruck oder aber die geringe Data-Ink-Ratio herangezogen werden. Um die Kernaussage hervorzuheben, muss die Entwicklung über die vergangenen Jahre verdeutlicht werden. Dies kann etwa durch eine lineare Interpolation geschehen, siehe dazu die Abbildung 2.11. Damit ist auch ein Beispiel für den zweiten grundlegenden Typ von Datenvisualisierung in der Ebene gegeben, der Linie. Einen deutlichen Nachteil bringt diese Abwandlung im vorliegenden Beispiel jedoch mit sich: Die diskreten Werte bekommen durch die Interpolation einen kontinuierlichen Charakter. Dieser Eindruck entspricht jedoch nicht der Wahrheit, denn der Bierabsatz schwankt innerhalb eines Jahres, selbst wenn man die Werte proportional auf Quartale oder Monate herunterrechnet. Um den täuschenden Effekt abzumildern, bleiben die diskreten Messpunkte in der Visualisierung enthalten.

Im Allgemeinen muss abgewogen werden, ob ein solcher Kompromiss der Verständlichkeit beiträgt, oder die Aussage der Visualisierung zu stark verfälscht. Als Alternative

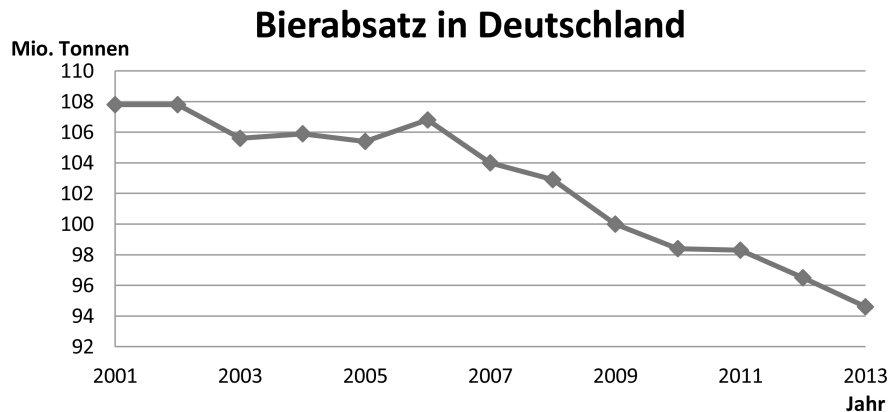


Abbildung 2.11.: Bierabsatz in Deutschland zwischen 2001 und 2013 unter Zuhilfenahme einer linearen Interpolation

Datenquelle: DESTATIS – ohne Autor (2014b, Tabelle)

Visualisierung: eigenes Werk

wurde in der Abbildung 2.12 auf eine Interpolation der Daten verzichtet und stattdessen ein linearer Trend eingezeichnet. Dieser hebt sich in der Symbolik von den Datenpunkten ab und vermeidet so Verwechslungen. Trotzdem birgt auch die Verwendung einer solchen Methode Risiken für die Kommunikation. So muss sich die Annahme eines gewissen Trends mit den eigentlichen Aussagen decken. Besonders kritisch ist die Voraussage einer Entwicklung unter einem Trend aus der Vergangenheit. Während für einen begrenzten historischen Zeitraum ein objektiver Trend angegeben werden kann, so hat dieser im Allgemeinen keinen Einfluss auf die mittel- oder langfristige Entwicklung. Insbesondere ist hier der exponentielle Trend zu nennen. Es ist sogar möglich, dass sich der Autor mit solchen Annahmen unbeabsichtigt selbst täuscht und in der Folge falsche Ergebnisse publiziert, obwohl das methodische Vorgehen ansonsten korrekt ist. Besondere Vorsicht ist in solchen Fällen bei einer explorativen Datenanalyse geboten. Die Aussagen sollten hier immer auf Gegenargumente geprüft werden, weil sie von der Visualisierung herrühren und nicht von einer vorformulierten Aussage.

Es bleibt der dritte Typ der visuellen Datenrepräsentation in der Ebene, die Fläche. Flächen drücken – ebenso wie Linien – in einem Diagramm einen kontinuierlichen Zusammenhang zwischen zwei Größen aus. So muss, um von der diskreten zur kontinuierlichen Darstellung zu gelangen, der Rand dieser Fläche ebenfalls interpoliert werden. Dies kann wie bei der Interpolation von Punkten zur Linie durch hinreichend genaue Messungen, durch theoretische Überlegungen oder Ähnlichem gerechtfertigt sein. Charakteristisch für die dargestellten Messgrößen ist es, dass die Werte der Abszisse und der Ordinate einen Einfluss auf die dargestellten Werte haben. Ein einfaches Beispiel eines solchen Zusammenhangs ist der Abbildung 2.13 zu entnehmen. Die Daten wurden entnommen und auf-

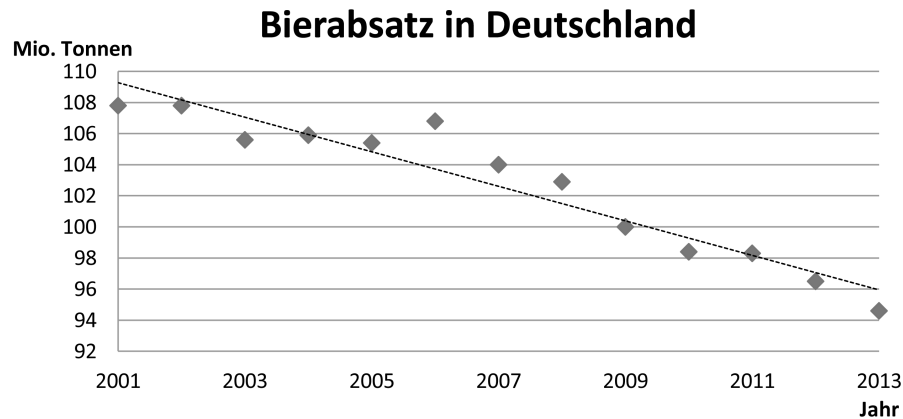


Abbildung 2.12.: Bierabsatz in Deutschland zwischen 2001 und 2013, Annahme eines linearen Trends

Datenquelle: DESTATIS – ohne Autor (2014b, Tabelle)

Visualisierung: eigenes Werk

bereitet nach einer Studienarbeit an der TU Clausthal⁶. Dargestellt ist die Wahrnehmung von Sprache durch das menschliche Gehör, die sowohl von der Frequenz (Abszisse), als auch der Lautstärke (Ordinate) eines Tons abhängt. Die dargestellte Kontur schließt den akustisch wahrnehmbaren Bereiche der beiden Komponenten im Zusammenspiel ein.

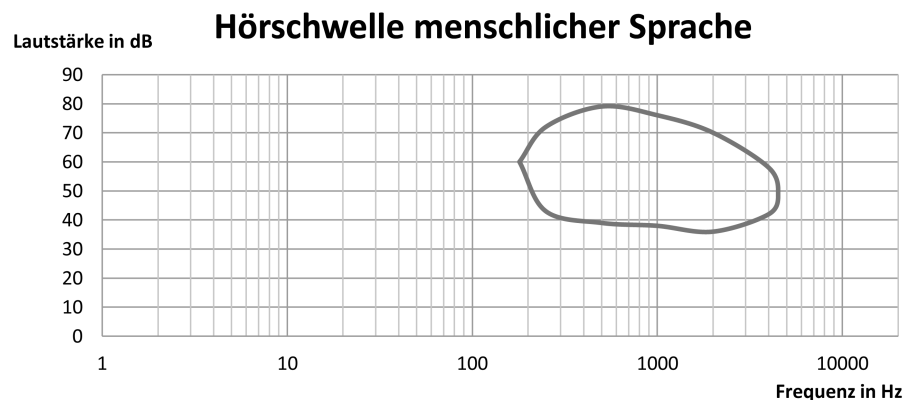


Abbildung 2.13.: Grenzen der menschlichen auditiven Wahrnehmung für Sprache

Datenquelle: Studienarbeit an der TU Clausthal⁶

Visualisierung: eigenes Werk

Neben den hier dargestellten quantitativen Komponenten können Diagramme auch Daten visualisieren, die *symbolische* Komponenten abbilden. Das sind alle Daten die keine

⁶Studienarbeit zum Thema Maschinenakustik

<http://www3.imw.tu-clausthal.de/forschung/projekte/EQUIP/studiarbeit/gehoer.html>

quantitative Ausprägung haben und dadurch nicht unmittelbar einer Ordnung unterliegen. Sie sind aber trotzdem komparativ, es ist also sinnvoll, diese Daten miteinander zu vergleichen. Solche Diagramme sind nicht mehr auf eine Darstellung in einem (kartesischen) Koordinatensystem angewiesen und können sehr unterschiedliche Formen annehmen. Hier spielt der künstlerisch-ästhetische Aspekt eine große Rolle, weshalb derartige Diagramme durch Ausschmückungen oft eine geringe Data-Ink-Ratio aufweisen. Auf weitere Beispiele dieser Art wird wegen der großen Verbreitung von Diagrammen zur Visualisierung verzichtet.

2.3.2. Zusätzliche Dimensionen in der Koordinatenebene

In den Beispielen des vorigen Abschnitts wurde jeweils lediglich ein Datensatz visualisiert. Dies führt dazu, dass größere Flächen im Diagramm leer bleiben. Dieser Raum steht also theoretisch weiteren Datensätzen zur Verfügung, durch deren Ergänzung sich auch die Data-Ink-Ratio erhöhen würde.

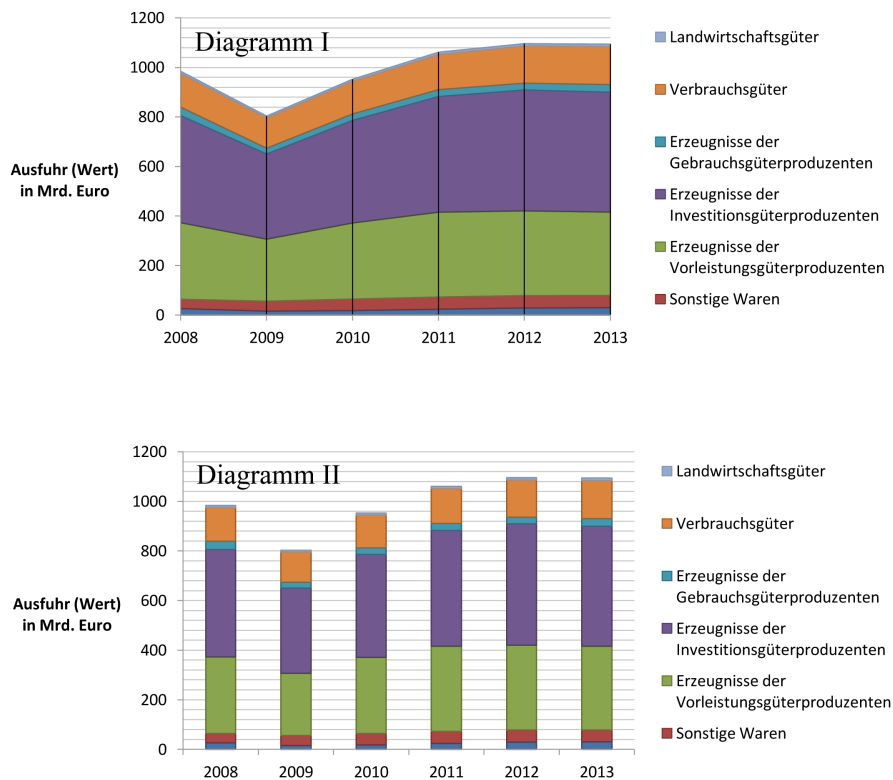


Abbildung 2.14.: Exporte der Bundesrepublik Deutschland der Jahre 2008 bis 2013, nach sieben Gütergruppen, in Euro.

Datenquelle: DESTATIS – ohne Autor (2014a)

Visualisierung: eigenes Werk

Ein Beispiel, welches mehrere Datensätze in einem einzigen Diagramm zeigt, ist in der Abbildung 2.14 zu sehen. Die Diagramme I und II bilden die gleichen Daten ab. Neben

den zwei Dimensionen in der Ebene wird eine dritte Dimension eingeführt. Abgebildet ist der Wert aller jährlichen Exporte der Bundesrepublik Deutschland in den Jahren von 2008 bis 2013. Zusätzlich ist der Gesamtexport in sieben verschiedene Gütergruppen unterteilt. Dadurch lässt sich einerseits eine Variation im Gesamtexport erkennen. Andererseits lässt sich durch die Aufteilung der Gütergruppen deren individuelle Variation und Beitrag zum Gesamtexport ablesen. Aus einer dritten Dimension lassen sich hier also Daten über insgesamt acht Messgrößen ableiten. Die Datendichte und die Data-Ink-Ratio werden entscheidend erhöht. Variiert wurde in beiden Diagrammen bezüglich der Gütergruppen das optische Attribut des Farbtons, da der Farbton keine natürliche Ordnung impliziert.

Nebenbei wurde von Diagramm I zu Diagramm II ein Optimierungsprozess durchgeführt. Beide Diagramme stellen exakt die gleichen Werte dar. Durch die Optimierung wird erstens von einer falschen Interpolation zwischen den Jahren abgesehen, zweitens werden die Anteile der Warengruppen deutlicher und drittens sind nun die Hilfslinien nahe jedem Wert erkennbar. Ein Vorgehen, wie es im Abschnitt 2.2.2 theoretisch beschrieben wurde. In der Praxis ist immer auf die individuellen Probleme einer Visualisierung einzugehen.

2.3.3. Die Landkarte als Metapher der Visualisierung

In der Einleitung zu diesem Abschnitt wurde bereits beschrieben, welchen Nutzen es hat, eine Visualisierung in eine Metapher einzubetten. Begründet wurde dies mit der intuitiven Verständlichkeit und dem ausgeprägten Gefühl für natürliche Bezugsgrößen. Eine besonders bekannte Metapher für Visualisierungen ist die Landkarte. Dabei handelt es sich in den häufigsten Fällen sogar um Darstellungen mit einer realen geografischen Landkarte als Basis. Das Konzept kann jedoch auch als relationale Landkarte ohne einen geografischen Bezug verwendet werden, ohne dass es der Wirkung abträglich wäre.

Die Verwendung einer Landkarte als Basis bringt mit sich, dass immer mindestens drei Datenkomponenten dargestellt werden. Mindestens zwei Datenkomponenten in der Ebene und mindestens eine auf der „Oberfläche“ der Landkarte. Eine Verwendung findet die Landkarten-Metapher etwa in der Abbildung 2.15, die ein Kataster sämtlicher Fälle von Krebs unter Männern in den Vereinigten Staaten zwischen 1950 und 1969 (Tuft, 1986, S. 16f.) zeigt. Die geografische Karte der Vereinigten Staaten ist dabei unterteilt in die einzelnen Countys. Für jedes County ist durch die Darstellung in einer Graustufe erfasst, wie sich die Anzahl von Krebsfällen im Verhältnis zu den gesamten Vereinigten Staaten im Mittel verhält. Die Bedeutung der Graustufen ist der Legende zu entnehmen. Die Wahl der optischen Variation des Kontrast ist passend, da es sich um eine quantitative und somit geordnete Größe handelt. Es stechen sowohl Countys heraus, die als einzige in ihrem Umkreis sehr hohe oder niedrige Krebsraten aufweisen, als auch Regionen von mehreren Countys, die leicht bis stark erhöhte Krebsraten aufweisen. Im Detail liegen rund um größere Städte und Industriezentren Bereiche, in denen die Zahl der Krebsfälle im Vergleich zum Landesmittel deutlich erhöht ist.

Ein weiteres Beispiel, dass jedoch nicht auf einer echten geografischen Karte basiert, ist die Karte der Galaxien, wie sie in der Abbildung 2.6 im Abschnitt 2.2.2 dargestellt ist. Es handelt sich außerdem um eine Karte mit sehr hoher Informationsdichte, die trotzdem

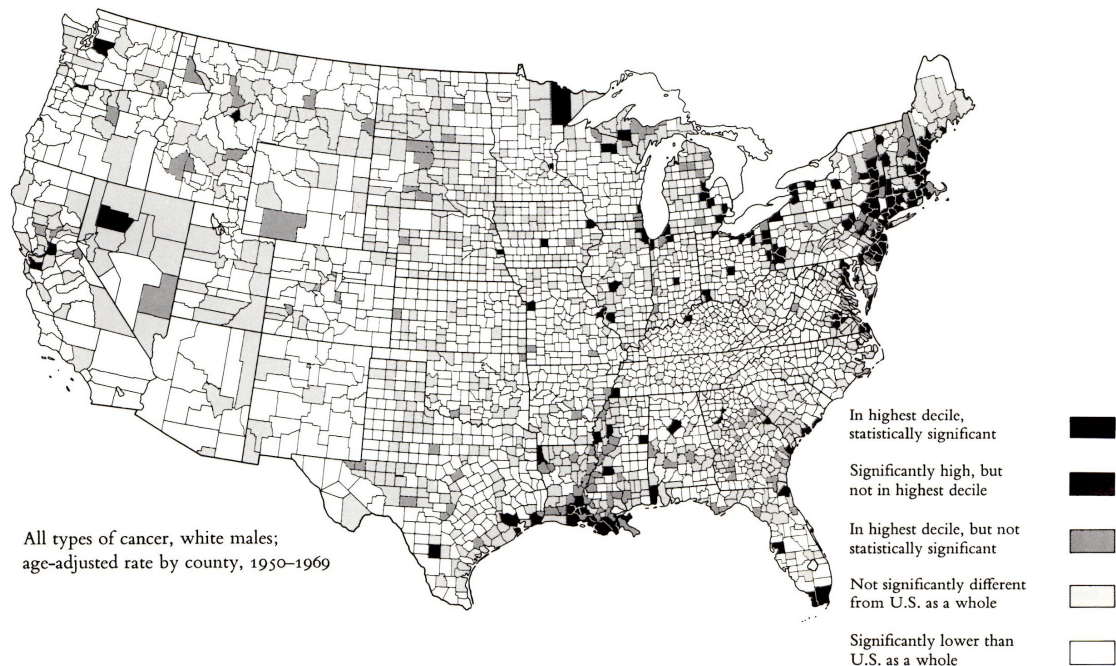


Abbildung 2.15.: Krebsfälle unter Männern pro County im Vergleich zu den gesamten Vereinigten Staaten, zwischen 1950 und 1969.

Quelle: Tufte (1986, S. 17)

leicht erfasst werden kann. Dies ist unter anderem wieder auf die Verwendung natürlicher Bezugsgrößen des Raumes zurückzuführen, obwohl diese in einem Wertebereich liegen, der im Alltag selten vorkommt. Der Wertebereich ist dort gar nicht genannt, sondern bezieht sich nur auf dimensionslose Quadrate, um den Betrachter nicht durch Zahlen von der Aussage der Visualisierung abzulenken (Tufte, 1986, S. 26).

2.3.4. Netzwerk-Visualisierungen mit Graphen

Ein Phänomen das in unserer Umwelt häufig anzutreffen ist, ist die Existenz von Systemen, die Einheiten enthalten und in denen diese Einheiten auf physikalische Weise oder aufgrund ihrer Eigenschaften verbunden sind. Physikalisch verbundene Netzwerke sind zum Beispiel Stromnetze, das Internet, Straßennetze oder das Nervensystem des Menschen. Auch nicht-vorhandene, aber theoretisch mögliche Verbindungen tragen zur Struktur eines Netzwerks bei. Systeme ohne physikalische Verbindung enthalten stattdessen virtuelle Verbindungen, die sich durch korrespondierende Eigenschaften der Einheiten auszeichnen. Beispiele hierfür sind soziale Kontakte, familiäre Beziehungen oder Zitationsbeziehungen zwischen wissenschaftlichen Publikationen.

Als formale Beschreibungssprache für Netzwerke kann die Graphnotation verwendet werden. Ein Graph kann Knoten und Kanten enthalten. Knoten repräsentieren Einheiten

im Netzwerk und Kanten die Verbindungen zwischen den Einheiten. Die Art der Kanten lässt sich in gerichtete und ungerichtete Kanten unterteilen. Bezogen auf das Beispiel der sozialen Kontakte sind die Verbindungen ungerichtet. Ein Beispiel ist der Abbildung 2.16 zu entnehmen: Kennt eine Person A eine andere Person B, so ist dies (zumindest im Allgemeinen) auch umgekehrt der Fall. Ein Beispiel für gerichtete Verbindungen ist in der Abbildung 2.17 zu sehen: Zitiert eine Publikation X eine andere Publikation Y, so impliziert diese Verbindung keinesfalls eine umgekehrte Zitation der Publikation X durch die Publikation Y.

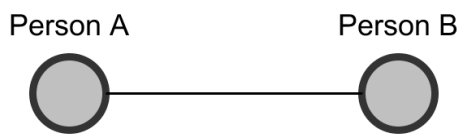


Abbildung 2.16.: Freundschaftsbeziehungen zwischen zwei Personen sind ungerichtet
Quelle: eigenes Werk

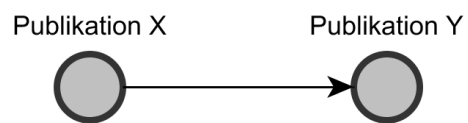


Abbildung 2.17.: Zitationsbeziehungen zwischen Publikationen sind gerichtet
Quelle: eigenes Werk



Abbildung 2.18.: Unabhängige Eigenschaft der Knoten: links männliches Geschlecht (blau) und rechts weibliches Geschlecht (rot). Abhängige Eigenschaft der Knoten: Anzahl der sozialen Beziehungen, in eckigen Klammern.
Quelle: eigenes Werk

Neben der reinen Existenz oder Nichtexistenz von Knoten und Kanten besitzen existierende Elemente weitere Eigenschaften. Diese Eigenschaften können sowohl abhängig vom Netzwerk als auch unabhängig davon sein. Wiederum angelehnt an das Beispiel der sozialen Kontakte kann jede Person die Eigenschaft männlich oder weiblich besitzen, welche unabhängig von den enthaltenen Verbindungen ist. Eine abhängige Eigenschaft wäre hier etwa die Anzahl der Freunde, denn diese wird durch die Anzahl der Verbindungen einer Person zu anderen Personen definiert. Die grafische Repräsentation ist der Abbildung 2.18 zu entnehmen.

Eigenschaften, die von der Struktur des Netzwerks abhängen, gegebenenfalls im Zusammenspiel mit unabhängigen Attributen, werden auch als Metriken bezeichnet. Besonders interessant sind jene Metriken, die einen weiteren semantischen Gehalt bieten. Unter der Annahme, dass die Anzahl der Freunde eine Aussagekraft über die Wichtigkeit der Person bezüglich einer Werbekampagne liefert, ließen sich mit Hilfe eines Netzwerks die besonders

wichtigen Personen leicht durch ihren Grad⁷ ermitteln. Als Rückschluss könnte nun die Werbekampagne entsprechend auf die Zielgruppe der „wichtigen“ Personen ausgerichtet werden.

Eine weitere Eigenschaft von (Teil-) Netzwerken ist die starke oder schwache Verknüpfung der Knoten untereinander durch Kanten. In einem Netzwerk existieren üblicherweise Mengen von Knoten, die untereinander sehr stark verbunden sind. Existieren mehrere dieser Mengen, sind die Knoten aus verschiedenen Mengen eher selten oder sogar gar nicht verbunden. Diese Struktur ist in vielen physikalischen oder virtuellen Netzwerken zu finden und wird als „Kleine-Welten-Phänomen“ bezeichnet. Siehe dazu das Beispiel in der Abbildung 2.19. Es sind drei farblich gekennzeichnete Mengen von Knoten enthalten, sogenannte Cluster. Der orange und der blaue Cluster sind nur über die Knoten mit den Nummern 3 und 5 verbunden, der grüne Cluster ist mit keinem weiteren Cluster verbunden. Aus dieser Aufarbeitung des Netzwerks durch Daten-Mining lassen sich weitere Informationen erhalten. Autarke Teilnetzwerke sind ebenso erkennbar wie wichtige „Brückenknoten“.

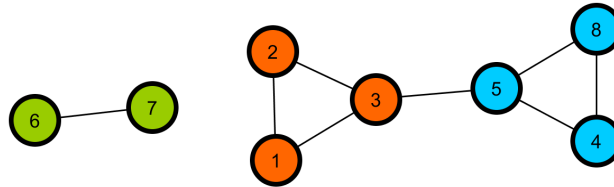


Abbildung 2.19.: Einteilung von Knoten in Cluster (jede Farbe entspricht einem Cluster), bezüglich ihrer Verbindungen.

Quelle: eigenes Werk

Anhand von Eigenschaften wie Metriken oder Clustern lassen sich Netzwerke auch visualisieren. Auf der einen Seite spielt hier die Anordnung der Knoten – und damit der Kanten – eine Rolle, auf der anderen Seite die Anreicherung durch Farbe, Form, Kontrast und andere visuelle Attribute. Diese Möglichkeiten machen die Netzwerk-Visualisierungen zu einem sehr mächtigen Werkzeug in der strukturellen Analyse von Daten. Die Voraussetzung ist hier lediglich, dass geeignete Konzepte zur semantischen Interpretation von Metriken, Clustern und der Anordnung von Knoten bestehen. So lassen sich in kurzer Zeit eine Vielzahl von komplexen Informationen über die strukturellen Eigenschaften der Daten ausdrücken.

2.3.5. Storyline Diagramme

Eine weitere Form der Visualisierung mit hoher Aussagekraft sind die sogenannten *Storyline-Diagramme*. Diese sind selten anzutreffen, da die Erstellung Fachkenntnis und Datensätze eines längeren Zeitraums erfordern. Das Erstellen von Hand bedarf vergleichsweise viel Aufwand. Ein Tool zur automatisierten Umsetzung solcher Storyline-Diagramme

⁷ Als Grad eines Knotens in einem Netzwerk wird die Anzahl der verbundenen Kanten bezeichnet.

bietet zum Beispiel das *MapEquation-Projekt* mit seinem *Alluvial Generator*. Die Storyline-Diagramme „erzählen eine Geschichte“ und bedienen sich dazu der natürlichen Bezugsgröße Zeit. Im Wandel der Zeit wird eine zweite Größe dargestellt, wobei sowohl die Zustände zu diskreten Zeitpunkten, als auch die Änderungen im Detail von Zeitpunkt zu Zeitpunkt deutlich werden. Zu jedem Zeitpunkt wird der gemessene Zustand eines Systems erfasst. Dieser wird in gerichteten oder ungerichteten Graphen mit Kanten- und Knotengewichten notiert. Das Tool bestimmt zwischen diesen Graphen die Änderungen und stellt sie als Fluss in einem an ein Histogramm angelehnten sogenannten *Alluvial Diagramm* dar. Die Darstellung zu den Zeitpunkten ähnelt einem Balkendiagramm, was den ursprünglichen Daten entspricht. Die Darstellung zwischen den Zeitpunkten ähnelt einem Flussdiagramm, dessen Informationen durch Daten-Mining aus den ursprünglichen Daten generiert wurden. Das aus den einzelnen Graphen resultierende Alluvial Diagramm vereint auf diese Weise mehrere Arten von Information in einer einzigen Visualisierung.

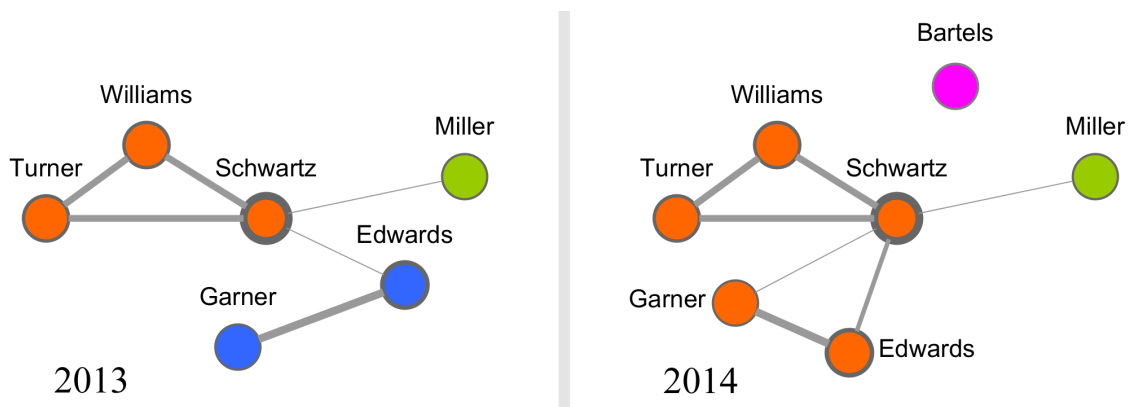


Abbildung 2.20.: Fiktive Zitationsbeziehungen zwischen Autoren. Zustand um 2013 (links) und 2014 (rechts), mit jeweiliger Einteilung in Cluster.

Quelle: eigenes Werk

Nachfolgend soll das vom MapEquation-Projekt verwendete Verfahren durch ein Beispiel verdeutlicht werden. In diesem Abschnitt liegt der Fokus auf dem verwendeten Verfahren, während im Kapitel 3 die eigentliche Funktionsweise des Alluvial Generators bezüglich der Aufarbeitung bibliografischer Daten dargestellt wird. Als Basis für das Beispiel dient eine Gruppe von sechs beziehungsweise sieben verschiedenen fiktiven Autoren. Diese haben sich in mehreren ihrer Arbeiten gegenseitig zitiert. Die Zitations-Beziehungen sind in der Abbildung 2.20 dargestellt. Zur Vereinfachung wurden ungerichtete Graphen verwendet. Die Breite der Kanten zwischen den Autoren repräsentiert die Anzahl der jeweiligen Zitationen, wobei eine breite Kante viele Zitationen repräsentiert. Die Breite der Kontur jedes Knotens repräsentiert die generelle Relevanz dieses Autors, unabhängig von der Anzahl der Zitationen. Nach einer Heuristik wurden verschiedene Cluster von Autoren identifiziert, die sich gegenseitig häufig zitieren. Die Zugehörigkeit zu einem Cluster ist durch je eine Farbe gekennzeichnet. Die Identifikation von Clustern kann aber auch durch den Alluvial Generator automatisiert werden. Die entsprechende Heuristik und ihre Randbe-

dingungen werden näher im Journalartikel *Mapping Change in Large Networks* (Rosvall und Bergstrom, 2010) erläutert.

Ein einzelner Graph mit Zitations-Beziehungen enthält zwar weitreichende Informationen, stellt letztlich aber nur einen einzelnen Zustand dar, der mit Kanten- und Knotengewichten versehen und darauf basierend in Cluster aufgeteilt wurde. Das eigentliche Ziel ist es, Unterschiede zwischen verschiedenen Graphen über die Zeit darzustellen. Übertragen auf das Beispiel liegen die Unterschiede in der Relevanz jedes einzelnen Clusters von Autoren und deren Entwicklung. Es wird also mindestens ein zweiter Graph benötigt, welcher entsprechende Veränderungen abbildet. Dazu sind in der Abbildung 2.20 zwei verschiedene Graphen enthalten, wobei der linke Graph den Zustand um 2013 und der rechte Graph den Zustand um 2014 darstellt. Die Veränderungen von 2013 auf 2014 sind im Wesentlichen folgende: Ein weiterer Autor namens *Bartels* ist hinzugekommen, er besitzt aber keine Zitations-Beziehungen zu anderen Autoren. Es bestehen außerdem noch zusätzliche Zitations-Beziehungen zwischen den Autoren *Edwards*, *Garner* und *Schwartz*, sodass diese nun nach der Heuristik dem selben Cluster angehören.

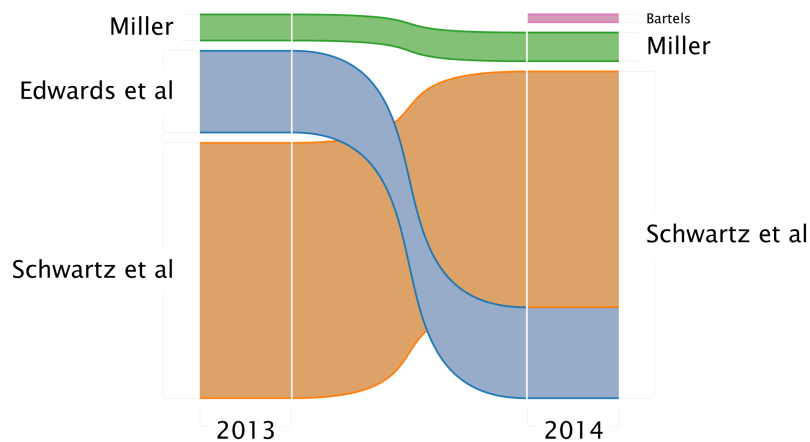


Abbildung 2.21.: Alluvial Diagramm der Zitations-Beziehungen zwischen sieben Autoren und der zeitliche Wandel der Beziehungen.

Quelle: Eigener Screenshot des Alluvial Generators mit dem vorliegenden Datensatz

Die abgebildeten Informationen lassen sich mit dem Alluvial Generator auswerten und im Alluvial Diagramm automatisiert veranschaulichen. In der Abbildung 2.21 ist das zu diesem Beispiel gehörende Alluvial Diagramm zu sehen. Die beschriebenen Änderungen zwischen den Zuständen um 2013 und 2014 werden so auf einen Blick kompakt visualisiert. Das gezeigte Beispiel bedient sich der Einfachheit halber lediglich zweier Zeitpunkte, der Fluss lässt sich aber über beliebig viele Zeitpunkte hinweg darstellen. Der Informationsgewinn durch diese Visualisierung ist, dass hervorgeht, wodurch eine wachsende oder sinkende Relevanz bezüglich der Zitations-Beziehungen zustande kommt. Voraussetzung ist natürlich, dass die Hypothese der Relevanz durch Zitations-Beziehungen korrekt ist.

Eine Visualisierung mit Hilfe von Storyline-Diagrammen ist für viele Arten von Informationen denkbar. Ebenfalls im Journalartikel *Mapping Change in Large Networks* stellen

die Autoren ein Alluvial-Diagramm vor, in dem sie die Entstehung der wissenschaftlichen Disziplin *Neuroscience* aus Teilen anderer Disziplinen über sechs Jahre dokumentieren (Rosvall und Bergstrom, 2010). Denkbar ist zum Beispiel auch eine Darstellung von „Hot-Topics“. Grundsätzlich lassen sich mit diesem Verfahren alle Information visualisieren, deren Topologie sich in einem Graphen ausdrücken lässt und die außerdem einem Wandel unterliegen.

2.3.6. Schlagwortwolken

Popularität haben in den letzten Jahren die sogenannten *Schlagwortwolken* (engl. Tag-cloud) erlangt. Viele Webseitenbetreiber setzen diese Form der Visualisierung an prominenter Stelle auf ihren Internetauftritten ein, um jedem Besucher in aller Kürze einen Überblick über den Inhalt des Internetauftritts zu geben. Da der Mensch beim Auffinden von Informationen in sehr kurzer Zeit in relevante und nicht-relevante Quellen unterteilt, kann eine Schlagwortwolke dabei helfen, die Aufmerksamkeit eines Webseitenbesuchers zu erlangen.



Abbildung 2.22.: Schlagwortwolken der berühmten Rede „I have a dream“ von Martin Luther King Jr., alphabetisch und zufällig angeordnet.

Idee nach Lohmann et al. (2009).

Quelle: Selbst erstellt mit TagCrowd <http://tagcrowd.com/> und Wordle <http://wordle.net/>.

Es gibt eine Vielzahl von verschiedenen Stilen, in denen Schlagwortwolken präsentiert werden können. Historisch wurde vielfach einzig und allein das Attribut der Schriftgröße variiert und die Begriffe in einer nicht-semantischen Ordnung ausgegeben. Die grafische Weiterentwicklung der Schlagwortwolken bringt aber eine Vielzahl an weiterer Information unter. So ist es mittlerweile üblich, die Anordnung der Schlagworte mit semantischem Gehalt zu versehen oder diese zumindest angenehmer für das menschliche Auge zu gestalten (Lohmann et al., 2009), siehe dazu die Schlagwortwolken in der Abbildung 2.22. Ein semantischer Gehalt kann zum Beispiel durch Anordnung nach gemeinsamem Auftreten der Schlagworte hinzugefügt werden. Schlagworte, die häufig zusammen verwendet werden,

sind räumlich näher positioniert. In der Folge bilden sich Cluster, die mit einem zusätzlichen visuellen Attribut, zum Beispiel einem Farbton, hervorgehoben werden können. Nach diesem Schema ist in der Abbildung 2.23 die Schlagwortwolke des gezeigten Beispiels in der Anordnung optimiert und durch Farbtöne in Cluster unterteilt.



Abbildung 2.23.: Schlagwortwolke der berühmten Rede „I have a dream“ von Martin Luther King Jr. in Clustern gruppiert bei gehäuftem gemeinsamen Auftreten.

Quelle: Lohmann et al. (2009)

Die Schlagwortwolke ist als Visualisierung durchaus auch für den praktischen Teil dieser Arbeit geeignet. Beispielsweise können als Neuinterpretation Co-Autorschaften und daraus folgende Autoren-Communities gut dargestellt werden, angelehnt an eine Darstellung wie in der Abbildung 2.23. Auch können relevante Schlagworte aus einem oder mehreren Texten, zum Beispiel allen Abstracts von Publikationen, gewonnen und dargestellt werden. So ließe sich etwa die thematische Ausrichtung in einer Schlagwortwolke darstellen.

2.4. Resümee und Einordnung

Die in diesem Kapitel aufgeführten Grundlagen der Visualisierung sind im Allgemeinen und speziell für den praktischen Teil dieser Arbeit geeignet, um verständliche und nützliche Visualisierungen zu erstellen. Ein Mehrwert durch die Verwendung von Visualisierungen als Hilfsmittel bei der Recherche von wissenschaftlicher Literatur ist klar gegeben. Die Aussagekraft einer einzelnen Publikation mit ihren Metadaten und dem Volltext ist begrenzt. Basis der wissenschaftlichen Arbeit ist es, Ergebnisse anderer Arbeiten zusammenzuführen, zu untermauern, zu widerlegen oder vertiefende Forschung zu betreiben. Alle diese Aufgaben stehen im Zusammenhang mit anderen Publikationen. Durch eine Einordnung einer Publikation in den Kontext anderer Publikationen wird ein Mehrwert der eigenen Ergebnisse erreicht.

Eine Organisation dieser Sammlungen, dem Kontext von Publikationen, ist herausfordernd und ohne eine digitale Literaturverwaltung heute kaum noch zu bewerkstelligen. Umso wichtiger ist es, auf Basis der dadurch entstehenden Möglichkeiten, die vielen gesammelten Daten noch besser aufzuarbeiten, als es traditionell üblich ist. Wie hier dargelegt wurde, können geeignete Visualisierungen diesem Ziel nützlich sein.

Diese Erkenntnis ist nicht neu oder allein Ergebnis dieser Arbeit. Viele vorangegangene theoretische und praktische Arbeiten haben sich bereits damit beschäftigt, in welcher Form Visualisierungen für eine Literaturrecherche oder die Aufarbeitung bibliografischer Sammlungen genutzt werden können. Um diese Ergebnisse auch für den praktischen Teil dieser Arbeit zu nutzen, widmet sich das folgende Kapitel 3 der Darstellung einiger praktischer Ansätze für diese Aufgaben. Ziel soll es auch sein, besonders wertvolle Visualisierungen zu identifizieren, um diese in die spätere Implementierung einfließen zu lassen.

3. Praktische Ansätze der bibliografischen Organisation und Visualisierung

Eine definierte Teilaufgabe der theoretischen Arbeit ist es, bereits existierende Lösungen für die in der Einleitung genannten Herausforderungen darzustellen. Dies ist einerseits sinnvoll, um eine Bestandsaufnahme des derzeit vorherrschenden Umgangs mit bibliografischen Informationen und den Arbeitsabläufen bei der Recherche wissenschaftlicher Literatur vorzunehmen. Andererseits können durch diese Darstellung gelungene Konzepte in die praktische Umsetzung von AcaVis übernommen und Fehler vermieden werden. Besonders wertvoll sind dafür Erkenntnisse der betrachteten Ansätze auf sinnvolle und nicht sinnvolle Funktionen, sowie die Umsetzung dieser Funktionen auf Basis von wissenschaftlichen Publikationen. Soweit Funktionen aus den vorgestellten praktischen Ansätzen übernommen werden, wird darauf im Kapitel 5 bei der Konzeption von AcaVis hingewiesen.

Ein Anspruch von AcaVis ist es, einen Benutzer in der Ausübung seiner wissenschaftlichen Recherche unterstützen. Die Basis einer Recherche bildet immer die Sammlung und Organisation von Publikationen oder Informationen im Allgemeinen. Ein konkretes Mittel zur Unterstützung der Recherche soll die grafische Visualisierung sein, wobei die besondere Eignung von Visualisierungen für diesen Zweck im Kapitel 2 motiviert wurde. Außerdem spielt bei der nachfolgenden Betrachtung eine Rolle, dass die Benutzeroberfläche von AcaVis webbasiert sein soll.

Die angeführten Aspekte legen eine Zweiteilung der in diesem Kapitel vorgestellten Ansätze nahe: Erstens die Ansätze zur Visualisierung, besonders im Hinblick auf die Visualisierung von bibliografischen Daten, und zweitens die Ansätze zur Recherche und zur bibliografischen Organisation. Die jeweilige Ausführung der Benutzeroberfläche – desktop- oder webbasiert – wird als Eigenschaft der vorgestellten Ansätze aufgeführt. Alle Angaben beziehen sich auf die genannte Programmversion oder, falls keine Informationen zur Version verfügbar sind, auf den Entwicklungsstand zum Zeitpunkt der Bewertung im Juli 2014. Angaben aus externen Quellen beziehen sich auf die dort verwendete Version, sind aber in jedem Fall auch relevant für die jeweils aktuellste Version mit dem Stand Juli 2014, falls die Version abweichend ist.

3.1. Ansätze zur Visualisierung

Zunächst werden einige Programme vorgestellt, die eine Visualisierung bibliografischer Daten ermöglichen. Es sei bereits angemerkt, dass die Benutzeroberflächen aller hier vorgestellten Programme, abgesehen vom *MapEquation Alluvial Generator*, desktopbasiert sind.

Außerdem ist anzumerken, dass das vorgestellte Programm *Cytoscape* im Abschnitt 3.1.2 zunächst keinen direkten Bezug zur Visualisierung bibliografischer Daten besitzt. Das Programm ist aber durch seine Fähigkeiten der Organisation von großen Datenmengen und der Umsetzung von Algorithmen der Graph-Analyse relevant. Außerdem wird Cytoscape durch den ebenfalls vorgestellten praktischen Ansatz *Sci²Tool* zur Visualisierung verwendet.

3.1.1. Science of Science Tool (Sci²Tool)

Evaluierte Programmversion Sci²Tool 1.1 Beta

Verfügbarkeit Kostenlos, aber eine Anmeldung ist notwendig

Benutzeroberfläche Desktopbasiert

Art der Visualisierung

Plots, Balkendiagramme, Netzwerke und Graphen, geografische Karten, 3D-Karten, auch Kombinationen aus diesen Visualisierungen

Weitere Funktionen

Statistische, zeitliche, thematische und geografische Analyse von Datensätzen. Insbesondere Analyse und Bereinigung von bibliografischen Informationen nach Publikationen, Autorschaft, Zitationen und Schlagworten. Basisfunktionen der Datenbereinigung für Listen, Netzwerke und geografische Daten. Analyse von Netzwerken und Volltextanalyse.

Einsatzgebiete

Sci²Tool ist geeignet für jede Disziplin welche die bereitgestellten Methoden zu nutzen vermag. Besondere Funktionen der Aufarbeitung bibliografischer Informationen wurden integriert. Allgemein liegt ein gewisser Fokus auf der Analyse von wissensbasierten Systemen. Es bestehen auch Online-Schnittstellen zu einigen sozialen Netzwerken.

Voraussetzungen für die Benutzung

Sci²Tool ist in der Programmiersprache Java geschrieben. Es basiert auf einer Eclipse-Equinox-Umgebung und bindet verschiedene andere Tools zur Verwendung über die Benutzeroberfläche ein. Es stehen Programmpakete für die Betriebssysteme Windows, MacOS und Linux bereit⁸.

Entwicklungsstand

Das Programm und dessen Erweiterungen werden weiterentwickelt.

Mitgelieferte Beispiele

Es werden Beispieldatensätze bei der Installation mitgeliefert. Deren Verwendung wird in der Dokumentation beschrieben.

⁸Sci²Tool Download <https://sci2.cns.iu.edu/user/download.php>

Erweiterbarkeit

Durch die Verwendung der Eclipse-Equinox-Oberfläche ist das Programm vollständig modular aufgebaut. Viele Funktionen sind prinzipiell durch externe Erweiterungen unter der Oberfläche von Sci²Tool realisiert. Die Java-Quelldateien sind Open Source⁹ und können aus dem zugehörigen Subversion-Repository¹⁰ ohne Anmeldung heruntergeladen werden. Anfragen für spezielle Erweiterungen können an das Entwicklerteam gestellt werden. Diese Erweiterungen werden gegebenenfalls auch gegen Bezahlung realisiert¹¹.

Benutzung und Funktionen

Die Benutzung des Programms basiert auf verschiedenen Arbeitsabläufen¹², welche die Module des Programms nacheinander in sinnvoller Weise einsetzen. Alle Arbeitsabläufe folgen einem ähnlichen Muster:

1. **Sammlung** von Rohdaten (Meistens nicht innerhalb von Sci²Tool)
2. **Verarbeitung** der Rohdaten für ein Datenmodell (Netzwerk, geografisch, ...)
3. **Bereinigung** der verarbeiteten Daten
4. **Analyse** von spezifischen Merkmalen (Insbesondere für Netzwerke)
5. **Visualisierung** der Daten durch ein passendes Verfahren

Zunächst müssen Daten erhoben werden, auf deren Grundlage eine Visualisierung stattfinden kann. In den meisten Fällen müssen diese manuell gesammelt werden. Sci²Tool ist in der Lage, diverse externe Formate zu verarbeiten. Darunter sind Netzwerkformate wie GraphML-Dateien und Pajek-Dateien, bibliografische Exportformate aus Datenbanken wie Scopus, Web of Science und Weiteren. Außerdem CSV-Dateien und einige weitere Formate mit speziellem Verwendungszweck. Eine umfängliche Auflistung der unterstützten Formate ist in der Dokumentation (Sci2 Team - ohne Autor, 2009) zu finden. Eine Besonderheit ist die Anbindung des Programms an einige soziale Netzwerke. Es bestehen Erweiterungen für Facebook, Flickr und Twitter. Diese ermöglichen es, Daten direkt im Programm online aus den Plattformen zu importieren und zu visualisieren. Nach Aussagen der Entwickler¹³ lässt sich so zum Beispiel der gesamte eigene Freundeskreis aus Facebook als Netzwerk visualisieren, sofern diese Daten freigegeben wurden.

Die eingelesenen Formate können entsprechend ihres Inhalts verarbeitet werden. Besonders viele solcher inhaltlichen Interpretation stehen für bibliografische Sammlungen zur Verfügung. Die in diesem Bereich gängigen Typen von Netzwerken wie Zitationen zwischen Autoren oder Publikationen, Co-Autorschaften, Schlagworte in Publikationen oder bibliografische Kopplung können automatisch extrahiert werden. Durch diese Möglichkeiten ergibt sich eine besondere Eignung des Programms für den Zweck der bibliografischen

⁹Lizenziert unter der Apache 2.0 Lizenz <http://www.apache.org/licenses/LICENSE-2.0>

¹⁰Subversion-Repository <https://in.cns.iu.edu/svn/nwb/>

¹¹Sci²Tool FAQ <http://wiki.cns.iu.edu/display/SCI2TUTORIAL/10+Sci2+FAQ>

¹²In der Dokumentation als „Workflow“ bezeichnet

¹³Integriert durch CI-Shell <http://wiki.cns.iu.edu/display/CISHELL/Facebook+Friends+Data>
CI-Shell ist als Plugin in das Programm eingebunden.

Analyse. Der Schritt kann weggelassen werden, wenn eingeleseene Daten bereits in strukturierter Form vorliegen.

Im dritten Schritt können die Daten auf vielerlei Weise bereinigt werden. Die Möglichkeiten sind jeweils individuell für den vorliegenden Datensatz als Netzwerk, geografischer Datensatz, Volltext oder als Liste zu wählen. Das Entfernen von isolierten Teilnetzwerken, von Zyklen oder die Vereinigung zweier Netzwerke ist möglich. Weitere Möglichkeiten bieten verschiedene Filtermechanismen. Relevant sind diese etwa für die Aufarbeitung von Postleitzahlen und Volltextanalyse. Für eine umfängliche Auflistung der Möglichkeiten sei auch hier wieder auf die Dokumentation (Sci2 Team - ohne Autor, 2009) verwiesen.

Im vierten Schritt, der Analyse, können durch Daten-Mining weitere Informationen gewonnen werden. Bezogen auf Netzwerke sind das beispielsweise Informationen über die kürzesten Pfade durch das Netzwerk, Knotengrade, PageRank der Knoten oder das Auffinden von Clustern im Netzwerk. Es können temporale und thematische Cluster auf Basis entsprechender Daten identifiziert werden. Auch ist es möglich, Daten mit dem in Sci²Tool integrierten *R-Softwarepaket*¹⁴ statistisch aufzuarbeiten.

Der letzte Schritt des Arbeitsablaufs führt zum eigentlichen Ziel, der Visualisierung. Die Algorithmen zur Darstellung der Visualisierungen werden von Sci²Tool lediglich aus anderen Programmen importiert. Es wird zum Beispiel auf die Programme Gephi¹⁵, Pajek¹⁶ und Cytoscape (wird im Abschnitt 3.1.2 behandelt) zurückgegriffen. Entsprechend der Ausrichtung des Programms sind insbesondere verschiedene Visualisierungen von Netzwerken und Graphen integriert. Je nach Wahl der Visualisierung können sämtliche optischen Attribute eines Netzwerks manuell oder in Abhängigkeit von Eigenschaften des Netzwerks dargestellt werden. Es gibt keine vorgegebene Belegung der optischen Attribute, sodass diese durch den Benutzer vorgenommen werden muss.

Beispiele

Auf eine Darstellung von konkreten Beispielen wird für Sci²Tool verzichtet. Die große Zahl an verfügbaren Visualisierungen erlaubt keine sinnvolle Auswahl von repräsentativen Kandidaten. Viele Möglichkeiten wurden bereits unter *Benutzung und Funktionen* theoretisch beschrieben. Umfangreiche Anleitungen und Beispiele sind außerdem in der Dokumentation (Sci2 Team - ohne Autor, 2009, Kapitel 5) enthalten, welche auch als Browser-Version¹⁷ verfügbar ist.

Bewertung

Sci²Tool liefert eine riesige Anzahl von Methoden zur visuellen Aufarbeitung von Daten. Der Fokus liegt dabei ganz besonders auf der Visualisierung von Netzwerken. Insbesondere sind viele spezielle Verfahren zur Aufarbeitung bibliografischer Datensätze enthalten. Diese

¹⁴The R Project for Statistical Computing <http://www.r-project.org/>

¹⁵Gephi-Projektseite <http://gephi.github.io/>

¹⁶Pajek-Projektseite <http://vlado.fmf.uni-lj.si/pub/networks/pajek/>

¹⁷Sci²Tool HTML-Dokumentation <http://wiki.cns.iu.edu/>

und die Vielzahl an weiteren Möglichkeiten zur Aufarbeitung machen das Programm zu einem umfangreichen und gelungenen Ansatz der praktischen Visualisierung, bezogen auf den Themenschwerpunkt dieser Arbeit.

Die Modularität des Programms ist eine seiner Stärken. Durch parallele Entwicklung und die Verwendung vieler bewährter externer Bibliotheken, welche die verschiedensten Teilprozesse im Arbeitsablauf abdecken, erhält Sci²Tool seine methodische Stärke. Besonders lobenswert ist der Open Source Ansatz, der es ermöglicht, eigene Erweiterungen zu erstellen und diese direkt in das vorhandene Umfeld einzubinden. Es entfällt dadurch viel eigener Entwicklungsaufwand, sodass gute individuelle Ergebnisse in akzeptabler Zeit erzielt werden können.

Jedoch ist ein Arbeiten mit dem Programm ohne Fachkenntnisse oder ohne eine geplante Vorgehensweise kaum möglich. Trotz der methodischen Stärke von Sci²Tool ist der Arbeitsablauf von der Datensammlung bis zur fertigen Visualisierung eine Herausforderung. Das Programm gibt den Verlauf einer Analyse nicht vor, sodass durch reines Ausprobieren der Funktionen in den seltensten Fällen sinnvolle Ergebnisse entstehen. Alleine das Sammeln und Bereinigen großer Datensätze bedarf eines tiefgreifenden Verständnisses der zugrundeliegenden Konzepte. Nach der eingängigen Benutzung des Programms komme ich deshalb zum folgenden Schluss: Für Benutzer, die das Vorgehen an sich beherrschen und lediglich eine Softwareunterstützung in ihrer Arbeit suchen, ist Sci²Tool die richtige Wahl. Zum spielerischen Erlernen von Techniken oder zum „Rumprobieren“ ist das Programm ungeeignet. Eine zu große Einarbeitungszeit ist nötig, die Ergebnisse erfüllen die Erwartungen sonst im Hinblick auf ihre Aussagekraft nur mit Einschränkungen.

3.1.2. Cytoscape

Evaluierte Programmversion Cytoscape 3.1.1

Verfügbarkeit Kostenlos, es ist keine Anmeldung notwendig

Benutzeroberfläche Desktopbasiert, eine Steuerung über die Konsole ist möglich

Art der Visualisierung Netzwerke und Graphen

Einsatzgebiete

Das Programm wurde nach Aussagen der Entwickler insbesondere für die Darstellung von Netzwerken in den Disziplinen Bioinformatik und Chemie entwickelt. Der Einsatz ist jedoch nicht darauf beschränkt.

Voraussetzungen für die Benutzung

Cytoscape ist in der Programmiersprache Java geschrieben und läuft in einer Java Virtual Machine der Versionen 6 und 7. Java 8 wird zu diesem Zeitpunkt noch nicht unterstützt¹⁸. Es stehen Programmpakete für die Betriebssysteme Windows, MacOS und Linux bereit. Die Cytoscape-Bibliotheken wurden auch als Javascript-Paket¹⁹ für die Benutzung auf Webseiten umgesetzt, dieses wird hier aber nicht betrachtet.

¹⁸Download und Anforderungen <http://www.cytoscape.org/download.php>

¹⁹Cytoscape in Javascript <http://cytoscape.github.io/cytoscape.js/>

Entwicklungsstand

Derzeit werden noch die Versionen 2.* und 3.* von Cytoscape verteilt. Die Verwendung der Version 3 wird empfohlen. Das Programm wird weiterentwickelt.

Mitgelieferte Beispiele

Einige Beispiele aus der Bioinformatik werden mitgeliefert. Deren Verwendung wird durch verschiedene Anleitungen²⁰ beschrieben, die über das Internet verfügbar sind.

Erweiterbarkeit

Über eine von den Entwicklern bereitgestellte Webseite²¹ können verschiedene geprüfte Erweiterungen in Cytoscape integriert werden. Auch die Integration von eigenen Erweiterungen wird unterstützt. Die Java-Quelldateien von Cytoscape sind Open Source²².

Benutzung

Die Funktionen bieten die Möglichkeit Beispielnetzwerke zu laden, Netzwerke zu modellieren oder Daten zu importieren. Die Importfunktion kann die verschiedensten gängigen Dateiformate für Graphnotationen verarbeiten. Darunter sind Excel-Tabellen, CSV-Dateien, das GML-Format und das GraphML-Format. Dabei kann entsprechend dem Verwendungszweck gewählt werden, welche Spalten der Quelldatei den Knoten und Kanten des Netzwerks zugeteilt werden sollen. Insbesondere muss angegeben werden, welche drei Spalten eines Datensatzes eine Kante genauer beschreiben. Die relevanten Eigenschaften einer Kante sind deren Start- und Zielknoten, sowie deren Zustand als gerichtete oder ungerichtete Kante. Außerdem existiert in Cytoscape ein Onlinemodul zur Datensammlung, welches jedoch in der Basisversion ausschließlich Informationen aus Protein-Datenbanken der Bioinformatik beziehen kann.

Nachdem ein Netzwerk geladen wurde, muss ein Layout für die Visualisierung des Graphen gewählt werden. Dabei ist das Erscheinungsbild unterteilt in ein Graphlayout, also die Anordnung der Knoten und ein optisches Layout, mit allen visuellen Attributen der Knoten und Kanten wie zum Beispiel Farbe, Form und Beschriftung. Für das Graphlayout sind verschiedene gängige Algorithmen implementiert. Neben einfachen hierarchischen, zirkulären oder baumartigen Graphlayouts stehen auch mehrere Verfahren zur Verfügung, die mit Hilfe von Knotengewichten ein kraftgerichtetes Layout umsetzen. So erhält die optische Nähe oder Distanz der Knoten eine Bedeutung bezüglich der Netzwerkstruktur. Das optische Layout ist vielfältig und kann in allen Punkten individuell angepasst werden. Einen besonderen Stellenwert erhält dieses dadurch, dass es mit den Attributen von Knoten und Kanten verknüpft werden kann. So können Knoten- und Kantengewichte durch die Größe der Knoten oder die Breite der Kanten dargestellt werden. Darüber hinaus können durch Farbe, Umrandung und Schriftgröße viele weitere Datenkomponenten in der Visualisierung untergebracht werden.

²⁰Cytoscape Tutorials http://opentutorials.cgl.ucsf.edu/index.php/Tutorial:Introduction_to_Cytoscape_3.1-part2

²¹Cytoscape Erweiterungen <http://apps.cytoscape.org/>

²²GPL 2.1 Lizenz <http://www.gnu.org/copyleft/lesser.html>

Die Informationen über Netzwerke und deren optische Layouts können separat oder gebündelt abgespeichert werden. Auf diese Weise lassen sich erstellte optische Layouts für verschiedene Netzwerke verwenden. Dadurch kann die geringe Zahl an vorgegebenen optischen Layouts kompensiert werden. Zusätzlich können die Visualisierungen in gängigen Vektorformaten und als Rastergrafik exportiert werden. Um einen Einblick in die Vielzahl an Möglichkeiten zu geben, die schon die Basisversion von Cytoscape bereitstellt, sind im Nachfolgenden einige Beispiele aufgeführt.

Beispiele

Es soll ein fiktives Beispiel die generelle Funktionsweise des optischen Layouts hervorheben. Der dargestellte Graph ist angelehnt an das Autoren-Netzwerk der Abbildung 2.20 aus Abschnitt 2.3.5. Die folgende Tabelle in der Abbildung 3.1 zeigt exakt die verwendeten Informationen. Diese könnten in der vorliegenden Form zum Beispiel mit Sci²Tool aufgearbeitet worden sein.

Entsprechend ergibt sich der Zitationsgraph wie in der Abbildung 3.2. Die Breite der gerichteten Kanten ist proportional zur Anzahl der Zitationen dargestellt. Der Durchmesser der Knoten ist proportional zur Anzahl aller Zitationen auf einen Autor dargestellt. Zum Beispiel hat *Miller* das Gewicht 3, weil dieser nur von *Schwartz* drei Mal zitiert wird.

Zitierender	Zitierter	Zitationen
Turner	Williams	5
Turner	Schwartz	5
Schwartz	Williams	10
Schwartz	Turner	12
Schwartz	Miller	3
Schwartz	Edwards	4
Williams	Turner	4
Williams	Schwartz	6
Edwards	Schwartz	2
Edwards	Garner	8
Garner	Edwards	7

Abbildung 3.1.: Zitationen zwischen sechs Autoren, fiktive Daten

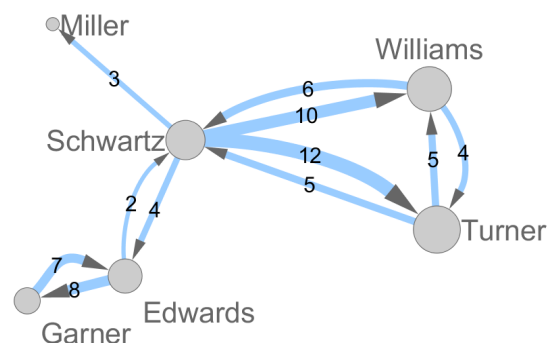


Abbildung 3.2.: Zitationsgraph zwischen sechs Autoren
Quelle: eigenes Werk

Bewertung

Obwohl die Zielgruppe von Cytoscape nach Angaben der Entwickler eher die Bioinformatik und Chemie sind, ist die Verwendung keinesfalls auf diese Bereiche begrenzt. Die Basisfunktionen sind bereits hinreichend geeignet für eine aussagekräftige und optisch ansprechende Visualisierung von Graphen jeder Art. Die allgemeine Verwendbarkeit wird auch belegt durch die Integration der Cytoscape-Programmpakete im bereits vorgestellten Sci²Tool. Die von der Cytoscape-Community bereitgestellten Erweiterungen verleihen

der Software einen Allround-Charakter. Neben der Visualisierung von Netzwerken können damit auch andere Formen der Visualisierung genutzt werden. Die Bandbreite der Erweiterungen reicht von gewöhnlichen Diagrammen über Cluster-Analyse in Graphen bis hin zu Importfunktionen für online verfügbare Daten. Die Funktionsweise von Cytoscape folgt einfachen Abläufen und ist so auch für Einsteiger geeignet. Tutorials und eine umfangreiche Dokumentation runden diesen Eindruck ab. Der hohe Grad an möglicher Individualisierung, begonnen bei den Quellformaten für Netzwerke, bis zur optischen Anpassung, ist vorbildlich. Zu bemerken ist aber, dass Cytoscape den Arbeitsspeicher stark belastet. Der Speicherverbrauch liegt mindestens im Bereich von 1,5 Gigabyte und steigt deutlich bei großen Netzwerken. Trotzdem läuft die Software aber stabil. Zusammengefasst ist Cytoscape ein sehr nützliches Programm für alle Visualisierungen rund um Graphen. Es ist außerdem gut für den Einstieg in die Visualisierungen und damit zusammenhängende Methoden geeignet.

3.1.3. CiteSpace

Evaluierte Programmversion CiteSpace 3.8 R5

Verfügbarkeit Kostenlos, es ist keine Anmeldung notwendig

Benutzeroberfläche Desktopbasiert

Art der Visualisierung

Hauptsächlich Netzwerke und Graphen. Diese werden erweitert durch Schlagwortwolken, die mit Clustern oder Knoten in Verbindung stehen. Außerdem können Netzwerke speziell an einer Zeitachse angeordnet werden.

Einsatzgebiete

CiteSpace analysiert bibliografische Datensätze. Dabei stehen die Volltextanalyse und die Zitationsanalyse im Vordergrund. Das Ziel von CiteSpace ist es, aus diesen Daten Informationen über ein wissenschaftliches Themenfeld, über Schlüsselpublikationen, die Forschungsaktivitäten auf einzelnen Gebieten und einschneidende Ereignisse identifizieren zu können, wie dessen Hauptentwickler Chaomei Chen anführt (Chen, 2006).

Voraussetzungen für die Benutzung

CiteSpace ist in der Programmiersprache Java geschrieben und ist optimiert für Java 7²³. Die Lauffähigkeit mit anderen Java-Versionen ist unbekannt. Das Programm wird als JAR-Archiv veröffentlicht und ist somit auf allen Betriebssystemen lauffähig, auf denen eine Java Laufzeitumgebung installiert ist.

Entwicklungsstand

Das Programm wird derzeit noch weiterentwickelt.

Mitgelieferte Beispiele

Ein Beispieldatensatz bezüglich Publikationen im Zusammenhang mit Terrorismus

²³CiteSpace Systemanforderungen <http://cluster.ischool.drexel.edu/~cchen/citespace/download.html>

in den vergangenen Jahrzehnten wird mitgeliefert. Die Verwendung von CiteSpace wird im Handbuch²⁴ und in kurzen Video-Anleitungen²⁵ beschrieben.

Erweiterbarkeit

Das Programm ist nicht erweiterbar, die Quelldateien werden nicht veröffentlicht.

Benutzung

CiteSpace ist in der Lage, auf einer Basis von bibliografischen Daten wichtige Eigenschaften dieser Daten hervorzuheben. Die Hauptfunktionalität dabei ist, Netzwerke zu visualisieren und diese hinsichtlich der Struktur zu analysieren. Zunächst muss also ein Datensatz mit bibliografischen Daten gesammelt werden. CiteSpace kann diese Daten entweder aus Textdateien importieren oder durch eine Online-Suche in der PubMed-Datenbank²⁶ für einen Suchbegriff generieren. Diese automatisierte Suche liefert derzeit aber keine Zitationen zu den Publikationen und ist somit nur für die Schlagwort-Analyse geeignet. Neben Zitationen lassen sich auch Autoren, Institutionen, Schlüsselwörter und Schlagworte aus Volltexten in Netzwerken visualisieren.

Die Datenbasis muss jeweils getrennt nach Jahren in einzelnen Textdateien geordnet sein. Für jedes Jahr wird mindestens eine Datei eingelesen, wobei im weiteren Verlauf eine Eingrenzung vorgenommen werden kann. Zur Verarbeitung der Rohdaten stehen weitere Filtermechanismen zur Verfügung. Es werden jeweils nur die obersten Elemente berücksichtigt, wobei die Anzahl variabel ist. Als Heuristik der Sortierung wird dabei die Häufigkeit verwendet, etwa bei einer Darstellung von Schlüsselwörtern als Knoten die Häufigkeit jedes einzelnen Schlüsselworts. Analog dazu wird mit anderen Daten verfahren. Die folgenden Einheiten aus der Datenbasis können als Knoten im Graph dargestellt werden: Autoren, Institutionen, Länder, Schlagworte, Disziplinen, Publikationen und Referenzen. Eine sinnvolle Auswahl muss durch den Benutzer getroffen werden. Die Visualisierung als Graph erfolgt entsprechend der Auswahl durch Co-Autorschaft, Co-Zitation, Co-Occurrence und Weiteren. Nach Abschluss der Bereinigung und der Auswahl der Daten steht die erste Visualisierung in Form eines ungerichteten Graphen zur Verfügung.

An der Visualisierung lassen sich verschiedene Modifikationen vornehmen. Das Netzwerk lässt sich durch Algorithmen in Cluster unterteilen. Cluster-Titel können automatisiert aus den Titeln der Publikationen, dem Abstract oder den hinterlegten Schlagworten der Publikationen erstellt werden. Die Auswahl der Begriffe für den Cluster-Titel kann durch die Verfahren *TF*IDF* oder *Log-Likelihood* erfolgen. Außerdem kann die Größe der enthaltenen Knoten durch verschiedene Maße beeinflusst werden. Zur Auswahl stehen das Zentralitätsmaß, die Eigenvektor-Metrik oder der PageRank eines Knotens. Welche Daten zur Analyse herangezogen werden, hängt von der Auswahl der Daten für die Knoten ab. Diverse optische Attribute können auch interaktiv eingestellt und unmittelbar als Änderung in der Visualisierung nachverfolgt werden.

Ein Abspeichern der vorgenommenen Änderungen an der Visualisierung ist möglich. Die zugrundeliegenden Netzwerk-Informationen können exportiert werden und es kann eine Rastergrafik des Netzwerks abgespeichert werden.

²⁴CiteSpace Handbuch <http://cluster.ischool.drexel.edu/~cchen/citespace/CiteSpaceManual.pdf>

²⁵CiteSpace Video-Anleitungen <http://cluster.ischool.drexel.edu/~cchen/citespace/videos/>

²⁶PubMed <http://www.ncbi.nlm.nih.gov/pubmed>

Beispiele

Im Folgenden wird zur Veranschaulichung ein Beispieldatensatz über Publikationen im Zusammenhang mit Terrorismus in den Jahren von 1999 bis 2003 verwendet. Dieser Datensatz ist zu Testzwecken bereits in CiteSpace enthalten und wurde ausführlich im Journalartikel *CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature* (Chen, 2006) besprochen. Die hier gezeigten Darstellungen wurden komplett neu erstellt und dienen der exemplarischen Veranschaulichung von CiteSpace, nicht der Aufarbeitung der Studie.

Der Datensatz enthält sämtliche bibliografische Daten über mehrere hundert Publikationen des Themengebiets, inklusive der Zitationen auf andere Publikationen. Als beispielhafter Arbeitsablauf wird die Visualisierung des zugehörigen *Document Co-Citation Graph* mit CiteSpace vorgenommen. Die entsprechende Darstellung ist in der Abbildung 3.3 zu sehen. Dargestellt sind als Knoten verschiedene Publikationen, die häufig in den Literaturverzeichnissen der Datenbasis enthalten waren. Eine Kante zwischen zwei Publikationen ist enthalten, wenn diese beiden Publikationen gemeinsam zitiert werden (Document Co-Citation). Die Breite der Kanten ist davon abhängig, wie oft diese zusammen genannt werden. In der hier abgebildeten Ansicht ist die Breite der Kanten zunächst irrelevant. Die Publikationen sind von CiteSpace auf Basis dieser Verbindungen in acht Cluster eingeteilt. Jeder Cluster ist mit dem häufigsten Schlagwort aller Publikations-Titel beschriftet. Als Sortierung der Schlagworte findet die Heuristik TF*IDF Anwendung.

Die beiden größten Cluster sind demnach *#0 medical response* und *#1 posttraumatic stress*. Die grobe thematische Verteilung der Publikationen lässt sich in dieser Darstellung erahnen. Je nach Ziel der Visualisierung können weitere Maßnahmen getroffen werden, um die Aussage zu verfeinern. Da hier keine echte Problemstellung zugrunde liegt, wird die Visualisierung nicht weiter optimiert.

In der nächsten Abbildung 3.4 wird nun der Cluster *posttraumatic stress* näher betrachtet. Die Kanten im Graph und die Beschriftungen der wichtigsten Publikationen sind nun deutlicher sichtbar. Jede Kante steht für gemeinsame Nennung zweier Publikationen in einem Literaturverzeichnis. Die Breite der Kante gibt, wie erwähnt, die jeweilige Häufigkeit an. Die Farbe jeder Kante gibt an, in welchem Jahr die Publikationen zuletzt gemeinsam genannt wurden. Die blauen Kanten im Jahr 1999, die gelben im Jahr 2002 und die orangen im Jahr 2003.

Die jeweiligen Publikationen lassen sich also zeitlich um die Terroranschläge vom 11. September 2001 in New York einordnen, da davon ausgegangen werden kann, dass das Phänomen von posttraumatischem Stress in den Jahren nach 2001 vermehrt Beachtung fand und hier ein Datensatz zum Thema Terrorismus zugrunde lag. Weitere Informationen lassen sich zunächst nicht entnehmen, da es sich wie erwähnt nicht um eine Visualisierung für eine konkrete Problemstellung handelt.

Bewertung

CiteSpace ist spezialisiert auf Problemstellungen bezüglich bibliografischer Sammlungen und thematischer Analyse. Aussagekräftige Visualisierungen diesbezüglich sind Graphen,

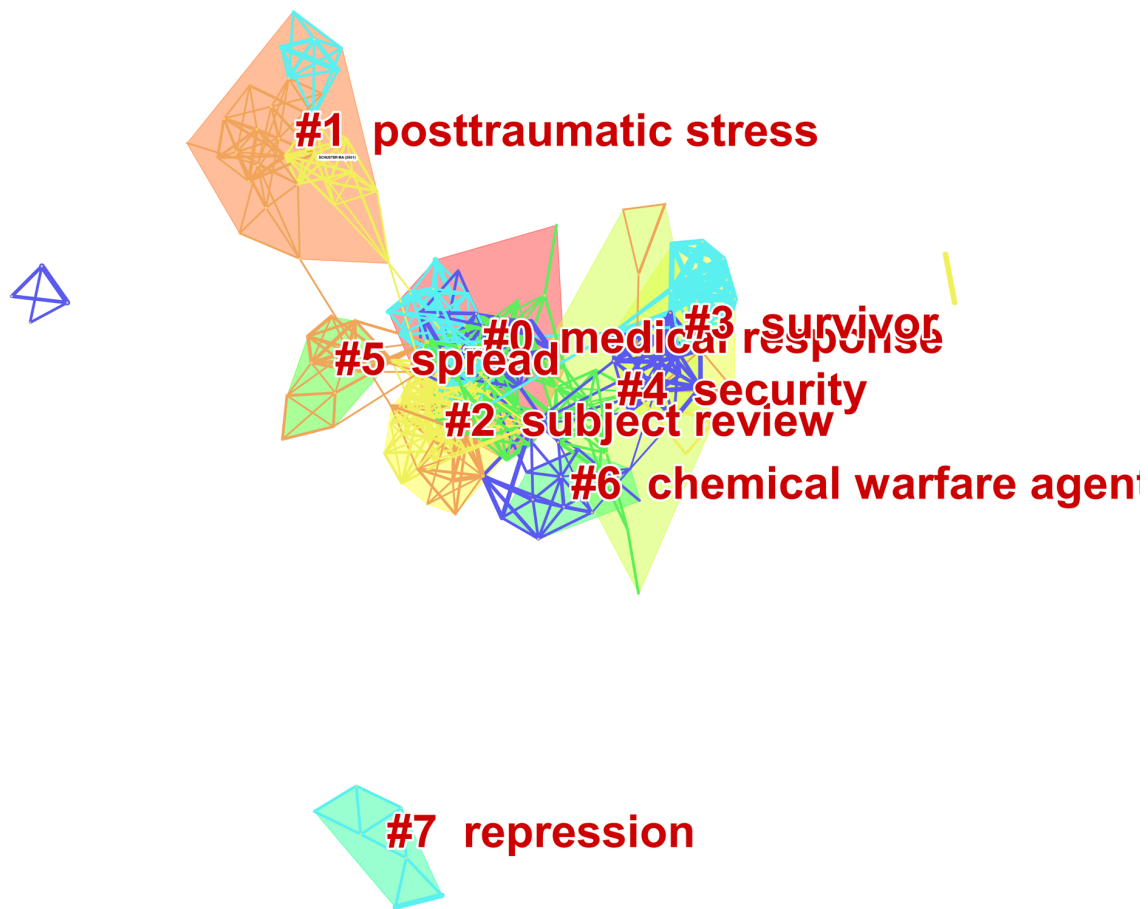


Abbildung 3.3.: Visualisierung von Publikationen aus dem Themenbereich Terrorismus der Jahre 1999 bis 2003
 Datenquelle: Chen (2006)
 Visualisierung: eigenes Werk, Screenshot aus CiteSpace

weshalb CiteSpace sich auf Graph-Visualisierungen spezialisiert. Die bereitgestellten Methoden zur Verarbeitung der bibliografischen Daten sind sehr umfangreich. Ein positiver Aspekt an CiteSpace ist die Anbindung an die bibliografische Datenbank PubMed, die es erlaubt, innerhalb des Programms Datensätze abzufragen. Leider ist die Anbindung derzeit für die meisten Analysemethoden in CiteSpace ungeeignet, da seitens PubMed häufig keine Daten mehr über Literaturverzeichnisse der Publikationen bereitgestellt werden. Die Benutzeroberfläche von CiteSpace ist unübersichtlich und suggeriert keinen klaren Arbeitsablauf. Die Lektüre der Anleitungen zu CiteSpace ist unerlässlich für eine sinnvolle Arbeit mit dem Programm. Dies führt dazu, dass umgekehrt ein exploratives oder „spielerisches“ Vorgehen schlecht möglich ist. Zu ähnlichen Ergebnissen kommen auch Allendoerfer et al. (2005) durch Nutzerstudien in ihrem Paper *Adapting the Cognitive Walkthrough Method to Assess the Usability of a Knowledge Domain Visualization*, wobei hier eine deutlich

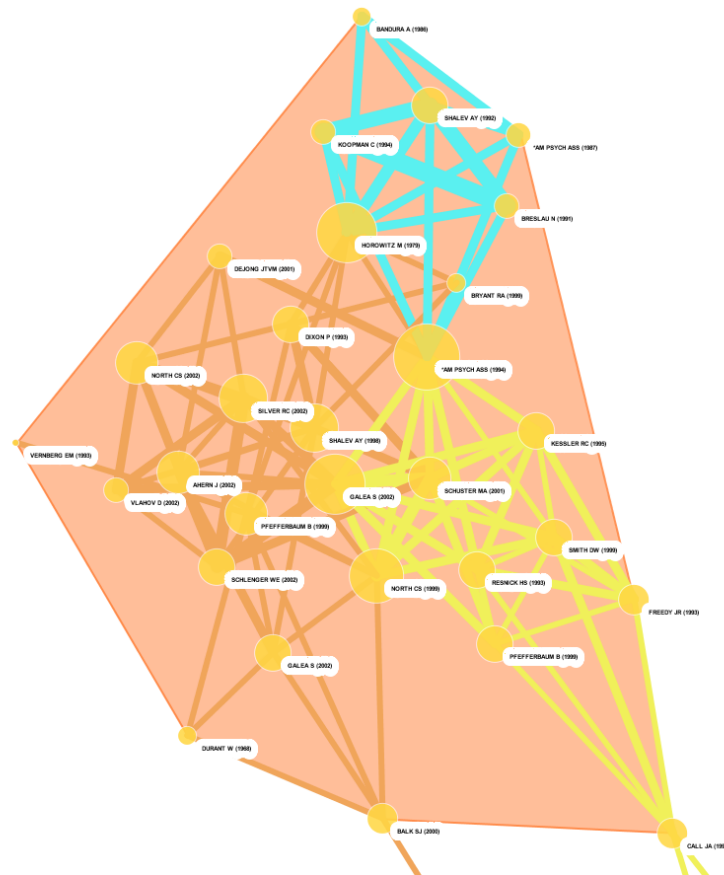


Abbildung 3.4.: Vergrößerte Ansicht des Clusters mit dem Titel *posttraumatic stress* aus der Abbildung 3.3

Datenquelle: Chen (2006)

Visualisierung: eigenes Werk, Screenshot aus CiteSpace

ältere Programmversion evaluiert wurde. Die Ergebnisse treffen erstaunlicherweise trotz der zeitlichen Differenz aber noch immer auf die aktuelle Version von CiteSpace zu. Die Performance des Programms ist verglichen mit modernen Applikationen relativ schlecht. Die Visualisierungen benötigen viel Zeit zur Darstellung. Die Interaktionen mit der Visualisierung sind daher nicht flüssig, der „Arbeitsfluss“ wird häufig durch Wartezeiten unterbrochen. Auch die Analysemethoden benötigen teilweise mehrere Minuten zur Berechnung, je nach Größe des Datensatzes. CiteSpace nutzt bei Mehrkernprozessoren nur einen Prozessorkern, was ebenfalls zur schlechten Performance beiträgt. Zusammengefasst ist CiteSpace also nur für Experten geeignet. Reich an Methoden lässt die praktische Implementierung aber einige Wünsche der Usability offen.

3.1.4. MapEquation Alluvial Generator

Evaluierte Programmversion Alluvial Generator Beta 0.3.6

Verfügbarkeit Kostenlos, es ist keine Anmeldung notwendig

Benutzeroberfläche Webbasiert, zugänglich über gängige Internetbrowser

Art der Visualisierung

Graphen, Berechnung von Clustern, Informationsflüsse zwischen Clustern und Entstehung von Clustern über die Zeit in Alluvial-Diagrammen.

Einsatzgebiete

Es gibt keine definierte Zielgruppe. Das Tool kann für sämtliche Anwendungen verwendet werden, in denen die Identifikation von Clustern und deren Vergleich über die Zeit sinnvoll erscheint. Zur Cluster-Berechnung wird der *MapEquation-Algorithmus* verwendet, beschrieben von Rosvall und Bergstrom (2008). Die Beispieldatensätze des Alluvial Generators beziehen sich insbesondere auf CoZitationen und die Darstellung wissenschaftlicher Themengebiete.

Voraussetzungen für die Benutzung

Der Map Generator und der Alluvial Generator stehen gemeinsam als Flash-Applikation über die Webseite des Projekts zur Verfügung. Voraussetzung für die Benutzung ist ein Webbrowser mit installiertem Flash-Plugin der Firma Adobe.

Entwicklungsstand

Unbekannt. Die letzten Änderungen liegen, gemessen an den veröffentlichten Publikationen zum Alluvial Generator, mehrere Jahre zurück.

Mitgelieferte Beispiele

Mehrere einzelne Beispieldatensätze über Co-Zitationen werden bereitgestellt. Ein weiterer Beispieldatensatz beinhaltet Informationen über wissenschaftliche Schlüsselbegriffe im Wandel über 6 Jahre.

Erweiterbarkeit

Das Programm ist nicht erweiterbar, die Quelldateien werden nicht veröffentlicht. Die zugrundeliegenden Algorithmen sind jedoch in den Publikationen der Autoren gut beschrieben und adaptierbar.

Benutzung

Zunächst muss die Webseite²⁷ des Tools über einen Internetbrowser aufgerufen werden, um die Flash-Applikation zu starten. Es steht dann die Eingabemaske für die Datenbasis zur Verfügung. Die Eingabeformate sind detailliert beschrieben und basieren auf denen des Programms Pajek.

Es gibt zwei grundsätzliche Formen von Dateiimport: Erstens können rohe Netzwerke mit Knoten und Kanten ohne Informationen über Cluster importiert werden, zweitens können bereits in Cluster unterteilte Netzwerke importiert werden. Im zweiten Fall kommt

²⁷<http://www.mapequation.org/apps/MapGenerator.html>

der MapEquation-Algorithmus nicht zur Anwendung. Im ersten Fall können die Cluster-Informationen entweder durch den MapEquation-Algorithmus generiert oder wiederum manuell vorgegeben werden. Für jedes Netzwerk müssen immer Clusterinformationen irgendeiner Art angegeben werden. Netzwerkinformationen werden im NET-Format, Clusterinformationen im CLU-Format und Kombinationen daraus im MAP-Format gespeichert. Die jeweilige Syntax ist der Projektseite²⁷ zu entnehmen. Die Netzwerke können auch direkt mit Pajek modelliert werden.

Sobald Netzwerk- und Clusterinformationen bereitgestellt wurden, können diese durch den Map Generator in einem Graph visualisiert werden. Das Layout des Graphen und das optische Layout lassen sich in vielerlei Hinsicht anpassen. Alle Graphlayouts berücksichtigen in unterschiedlicher Art und Weise die Knoten- und Kantengewichte.

Es können mehrere Netzwerke gleichzeitig bereitgestellt und in einer Zeitleiste angeordnet werden. Dies geschieht jeweils nach dem bereits beschriebenen Vorgehen. Sind nun mindestens zwei Netzwerke mit zugehörigen Clustern importiert, so lassen sich die Unterschiede zwischen je zwei Cluster-Netzwerken in einem Alluvial-Diagramm darstellen. Als identifizierendes Attribut für die Cluster dient deren Name. Die Entwicklung der Größe von Clustern wird im Alluvial-Diagramm ebenso visualisiert, wie die Zu- und Abflüsse von Informationen zwischen Clustern. Beliebige viele Netzwerke lassen sich in der Zeitleiste darstellen.

Eine schematische Darstellung zweier Netzwerke, unterteilt in Cluster, ist in der Abbildung 3.5 zu sehen. Die jeweiligen Cluster sind durch verschiedene Farben gekennzeichnet. Ein hoher Kontrast repräsentiert eine hohe Signifikanz der Zuordnung zu diesem Cluster. Das entsprechende Verfahren zur Identifikation von Clustern ist im Journalartikel *Mapping Change in Large Networks* (Rosvall und Bergstrom, 2010) beschrieben. Diese Unterteilung in Cluster unterliegt durch Änderungen im Netzwerk einem zeitlichen Wandel. Durch eine Änderung der *Flüsse* im Netzwerk, also der Kanten und deren Gewichte, ändert sich die Zuordnung einzelner Knoten zu Clustern. Als Flüsse werden die Verbindungen zwischen verschiedenen Clustern bezeichnet. Die Knoten ändern sich hier nicht. Die entsprechenden Unterschiede zeigen die Netzwerkgraphen in der Abbildung 3.5. Aus den strukturellen Änderungen im Netzwerk wurde das in der Abbildung 3.5 unten gezeigte Alluvial-Diagramm generiert. Von besonderem Interesse sind hier Zusammenflüsse und Divergenzen einzelner Cluster.

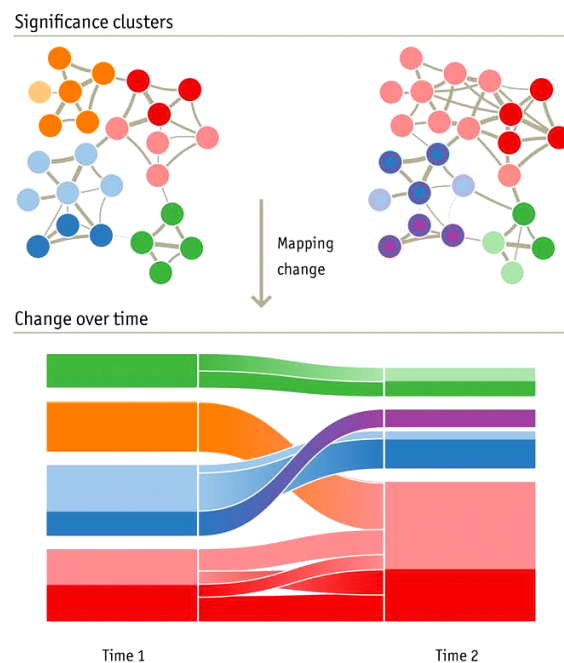


Abbildung 3.5.: Schematische Darstellung des Cluster-Wandels durch ein Alluvial-Diagramm. Quelle: Rosvall und Bergstrom (2010)

Interaktiv lassen sich auf der Webseite zudem die einzelnen Cluster auf ihre Flüsse hin untersuchen, da diese im globalen Diagramm bei großen Netzwerken schwer zu erkennen sind. Rosvall und Bergstrom (2008) beschreiben auch, dass die Betrachtung der Flüsse gegenüber der Betrachtung von ungerichteten Verbindungen methodische Vorteile hat. So ist der gesamte Informationsgehalt einer solchen Darstellung kaum in einer einzigen Grafik unterzubringen. Mit Hilfe der möglichen Interaktionen lassen sich die einzelnen Aspekte je nach Bedarf visualisieren. Die folgenden Beispiele stellen deshalb nur einige wichtige Besonderheiten heraus und liefern keine umfassende Erklärung der Visualisierungen. Weiterführende Erklärungen sind den angegebenen Quellen zu entnehmen.

Beispiele

Wie beschrieben wird zunächst ein Cluster-Graph des Netzwerks mit dem Map Generator visualisiert. Ein Beispiel eines solchen Cluster-Graphen ist in der Abbildung 3.6 zu sehen. Dieses enthält als Knoten insgesamt 51 verschiedene wissenschaftliche Disziplinen. Dabei handelt es sich um Cluster von Journalen, also den Knoten aus dem Ursprungsnetzwerk. Sie sind verknüpft über Kanten, deren Gewicht durch über 6 Millionen Zitationen zwischen Journalartikeln in insgesamt 6.128 Journalen bestimmt wird (Rosvall und Bergstrom, 2008). Diese Journale sind entsprechend des MapEquation-Algorithmus in Cluster eingeteilt, wie Rosvall und Bergstrom (2008) in ihrem Journalartikel *Maps of random walks on complex networks reveal community structure* anführen. Jeder dieser Cluster ist in der Abbildung 3.6 als Knoten dargestellt. Die Größe eines Knotens in der Darstellung ergibt sich aus dem summierten PageRank aller in diesem Cluster enthaltenen Knoten. Die gerichteten Kanten stellen einen Fluss von Zitationen von einem Cluster in einen anderen Cluster dar. Die Breite dieser Kanten ergibt sich aus den Flüssen, also den Zitationen entsprechend der eingeteilten Cluster (Rosvall und Bergstrom, 2008). Jeder einzelne Fluss wird relativ zum gesamten Fluss im Netzwerk betrachtet.

Basierend auf dem MapEquation-Algorithmus wird das Netzwerk zufällig durchlaufen und durch die Häufigkeit der aufgefundenen Disziplinen eine Unterteilung in Cluster vorgenommen. Nach Rosvall und Bergstrom (2008) fanden zum Beispiel 26 Prozent aller zufälligen Bewegungen im Netzwerk innerhalb der Disziplinen der Molekular- und Zellbiologie statt, diese Journale machen also 26 Prozent des gesamten Netzwerks aus. In der gezeigten Abbildung sind lediglich 98 Prozent aller Knoten und 94 Prozent aller Kanten enthalten, die Knoten mit dem kleinsten Gewicht und die Kanten mit dem kleinsten Fluss wurden durch Filterung entfernt (Rosvall und Bergstrom, 2010).

Die Autoren beschreiben auch einige wichtige Maßnahmen die ergriffen wurden, um die Aussagekraft der Visualisierung zu stärken. Einige Journale mit vielen interdisziplinären Artikeln wurden ignoriert, da diese eine relevante Verbindung von Disziplinen suggerieren, die real nicht existiert. Es wurden nur Journale mit mindestens zwölf veröffentlichten Artikeln pro Jahr berücksichtigt. Außerdem wurden nur Zitationen auf andere Journalartikel berücksichtigt, nicht auf sonstige Publikationen und auch keine Zitationen auf Artikel des selben Journals. Berücksichtigt wurden ausschließlich Journalartikel aus dem Jahr 2004 und deren Zitationen auf Journalartikel der fünf vorangegangenen Jahre. Die identifizierten Cluster wurden von den Autoren subjektiv anhand der enthaltenen Journale nach

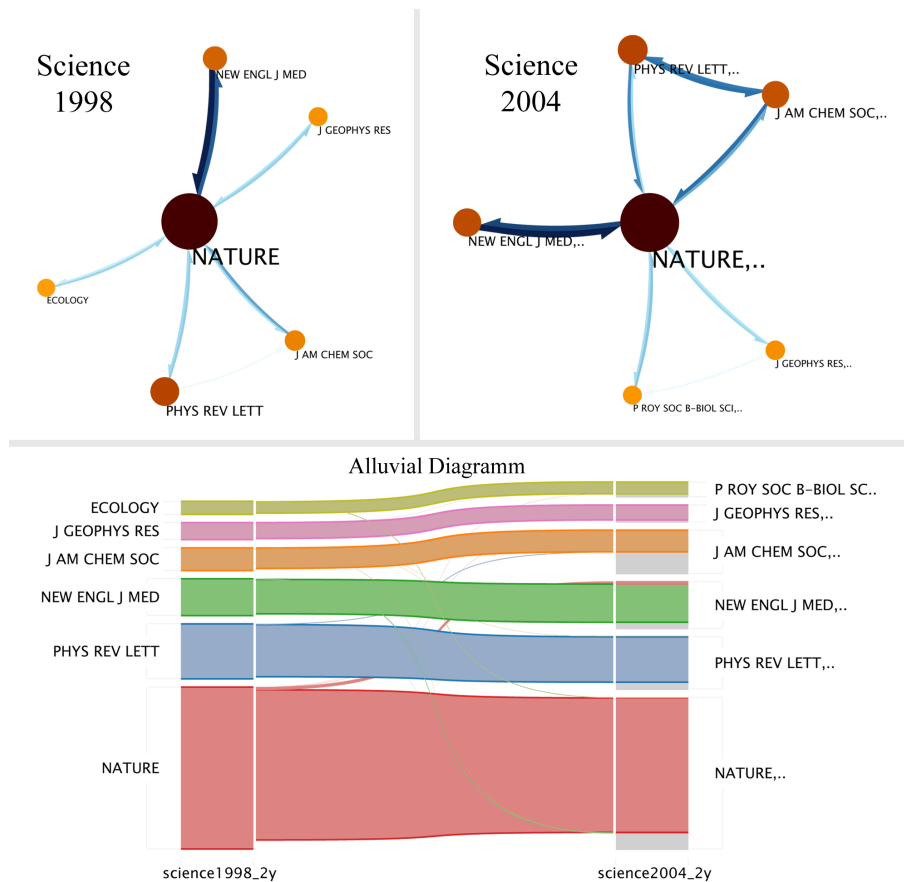


Abbildung 3.7.: Entstehung eines Alluvial Diagramms aus Cluster-Netzwerken
 Datenquelle: Beispieldaten des Alluvial Generators
 Visualisierung: eigenes Werk, Screenshot des Alluvial Generators

Clustern zwischen 1998 und 2004. Zum Beispiel ist zu sehen, dass einige Journale des roten Clusters im Jahr 2004 deutlich häufiger als zuvor Journale aus dem grünen Cluster zitieren. Alle grauen Abschnitte auf der rechten Seite des Alluvial-Diagramms sind Journale, die 1998 noch nicht Teil der Betrachtung waren. Der beschriebene Effekt ist zwar auch im Vergleich der Zitationen in den Map-Diagrammen zu sehen, wird jedoch nicht so deutlich wie im Alluvial-Diagramm. Insbesondere wenn es sich um große Netzwerke handelt, gestaltet sich ein direkter Vergleich schwierig. Um den Fokus auf die Funktionsweise des Alluvial-Diagramms und der Aussage zu lenken, wurde die Anzahl der dargestellten Cluster und Flüsse zwischen ihnen stark reduziert. Die Visualisierung stellt somit kein umfassendes Bild der Unterschiede zwischen 1998 und 2004 dar.

Bewertung

Der MapEquation Alluvial Generator bietet einen hervorragenden Ansatz der Visualisierung großer Datenmengen mit Netzwerk-Charakter. Der Arbeitsablauf ist klar strukturiert und nahezu selbsterklärend. Die Zugänglichkeit als Weboberfläche ist besonders zu begrüßen, auch wenn zur Darstellung das Adobe Flash-Plugin notwendig ist. Neben der reinen Darstellung von Netzwerken, die eine Vielzahl an methodischen Details enthält, wurde mit den Alluvial-Diagrammen ein relativ wenig beachteter Ansatz der Visualisierung implementiert. Wegen der starken optischen Aussagekraft dieser Flussdiagramme ist es verwunderlich, dass diese bisher noch nicht in breitem Umfang genutzt werden. Der MapEquation Alluvial Generator bedient sich der verbreiteten Pajek-Formate als Datenstruktur. Das ist einerseits von Nachteil, da von den Autoren insbesondere die Visualisierung von bibliografischen Daten als Anwendung vorgeschlagen wird, diese aber erst entsprechend in die benötigten Formate umgewandelt werden müssen. Andererseits ist diese Umsetzung auch von Vorteil, da es die Benutzung des MapEquation Alluvial Generators zusammen mit anderen Programmen erlaubt, die ihrerseits auf Pajek-Formaten aufbauen. Insgesamt ist das Programm also für die Aufgaben für die es konzipiert wurde gut umgesetzt.

3.1.5. Weitere Systeme zur Visualisierung

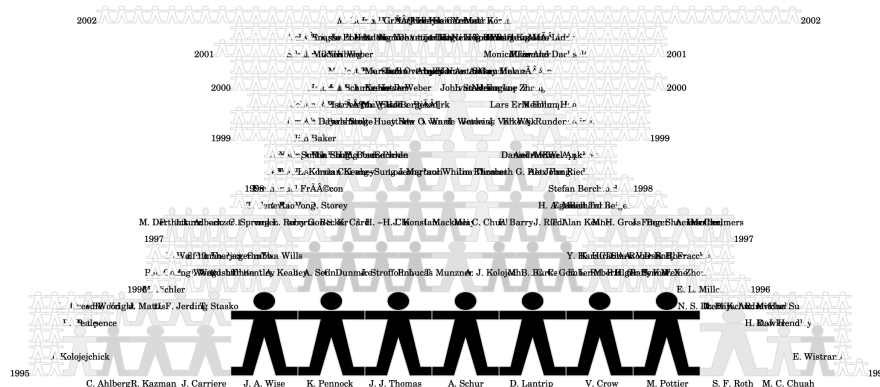
Neben den ausführlich vorgestellten Programmen existieren weitere Ansätze, die entweder nur einen geringen Funktionsumfang bieten, bereits eingestellt sind, oder aber nicht zur Evaluation zur Verfügung standen. Trotzdem ist es sinnvoll einige davon zu betrachten, um gelungene Ansätze zu adaptieren und Anregungen zu sammeln.

CiteWiz

CiteWiz wurde in den Jahren 2003 bis 2007 von Niklas Elmqvist und Philippas Tsigas als Prototyp an der Universität Göteborg entwickelt²⁸. Es handelt sich um eine Desktopanwendung in der Programmiersprache C++. Das Hauptziel des Programms sollte es sein, die Exploration wissenschaftlicher Publikationen zu erforschen und zu unterstützen (Elmqvist und Tsigas, 2007). Einerseits basierend auf traditioneller Suche innerhalb von bibliografischen Datenbanken, andererseits durch visualisierende Techniken. Dabei haben sich die Autoren an bestehenden theoretischen Überlegungen zu einem solchen Informationssystem bedient und diese dann prototypisch umgesetzt (Elmqvist und Tsigas, 2007). Es wurde auch eine Evaluation anhand von Nutzerstudien vorgenommen. Die Motivation, das Vorgehen und die Evaluation der Ergebnisse wurden im Journalartikel *CiteWiz: a tool for the visualization of scientific networks* von Elmqvist und Tsigas (2007) vorgestellt. Das Programm CiteWiz ist nicht öffentlich erhältlich. Daher beziehen sich die konkreten Informationen über das Programm im Nachfolgenden vollständig auf diese Publikation. Die gezogenen Schlüsse sind davon ausgenommen.

²⁸CiteWiz-Webseite <https://engineering.purdue.edu/~elm/projects/citewiz.html>

-



wird durch das Piktogramm eines Menschen dargestellt und erscheint im Jahr seines ersten Beitrags zu dieser Konferenz. Die Größe eines Autors ist proportional zu der Anzahl der Zitationen auf seine Publikationen. Es ist außerdem möglich, die Visualisierung zu vergrößern und so den Fokus auf Autoren mit weniger Einfluss zu lenken. Der Vorteil an dieser Darstellung sei es Elmqvist und Tsigas zufolge, dass Autoren mit besonderem Einfluss unmittelbar identifiziert werden können. Eine ähnliche Form der Visualisierung wurde auch für Publikationen entwickelt, die entsprechend chronologisch nach Erscheinungsjahr und in der Größe proportional zur Anzahl der Zitationen dargestellt werden.

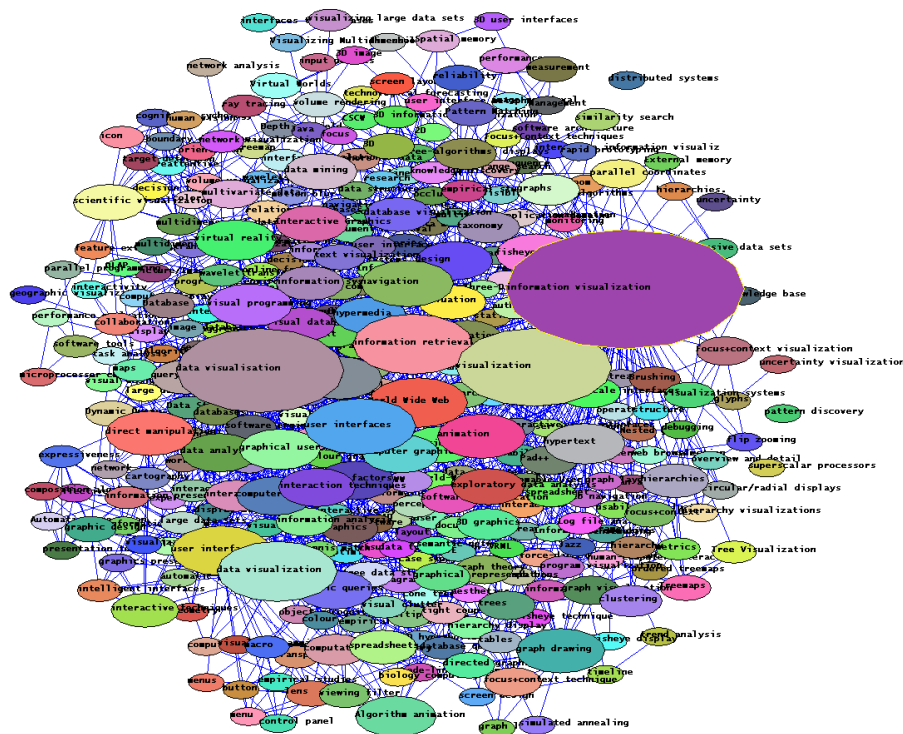


Abbildung 3.9.: Eine Graph-Visualisierung aus CiteWiz
Quelle: Elmqvist und Tsigas (2007)

Auch realisiert CiteWiz eine Graph-Visualisierung, die verschiedene Arten von Daten visualisieren kann. Co-Autorschaften, Zitationen und korrelierende Schlagworte sind einige davon (Elmqvist und Tsigas, 2007). In der Abbildung 3.9 ist ein solcher Graph von Schlagworten dargestellt. Die Bedeutung der Visualisierung beschreiben Elmqvist und Tsigas folgendermaßen: Jedes Schlagwort wird durch einen Knoten im Graph repräsentiert. Die Größe der Knoten wird durch die Häufigkeit des Schlagwortes bestimmt. Kanten zwischen den Knoten werden erstellt, falls diese in der selben Publikation zusammen enthalten sind. Das Layout des Graphen wird durch einen physikalischen Algorithmus berechnet, der das Knotengewicht – das ist wie angeführt die Häufigkeit eines Schlagwortes – berücksichtigt. Verschiedene Parameter erlauben es dem Benutzer, in sämtlichen Graphen die Anzahl der dargestellten Knoten beziehungsweise die minimale Häufigkeit einzustellen, sodass die

Übersichtlichkeit der Visualisierung gewahrt wird (Elmqvist und Tsigas, 2007).

Die histogrammische Visualisierung und die Graph-Visualisierung ermöglichen aus ihrer Natur heraus eher eine grobe Navigation im vorliegenden Datensatz. Für eine detailliertere Betrachtung wird ein dritter Typ der Visualisierung verwendet, dieser nennt sich *Influence-Visualisierung*. Ein Beispiel für eine solche Visualisierung ist in der Abbildung 3.10 zu sehen. Eine Gruppe von Autoren wird hier hinsichtlich ihrer gegenseitigen Zitationen näher betrachtet.

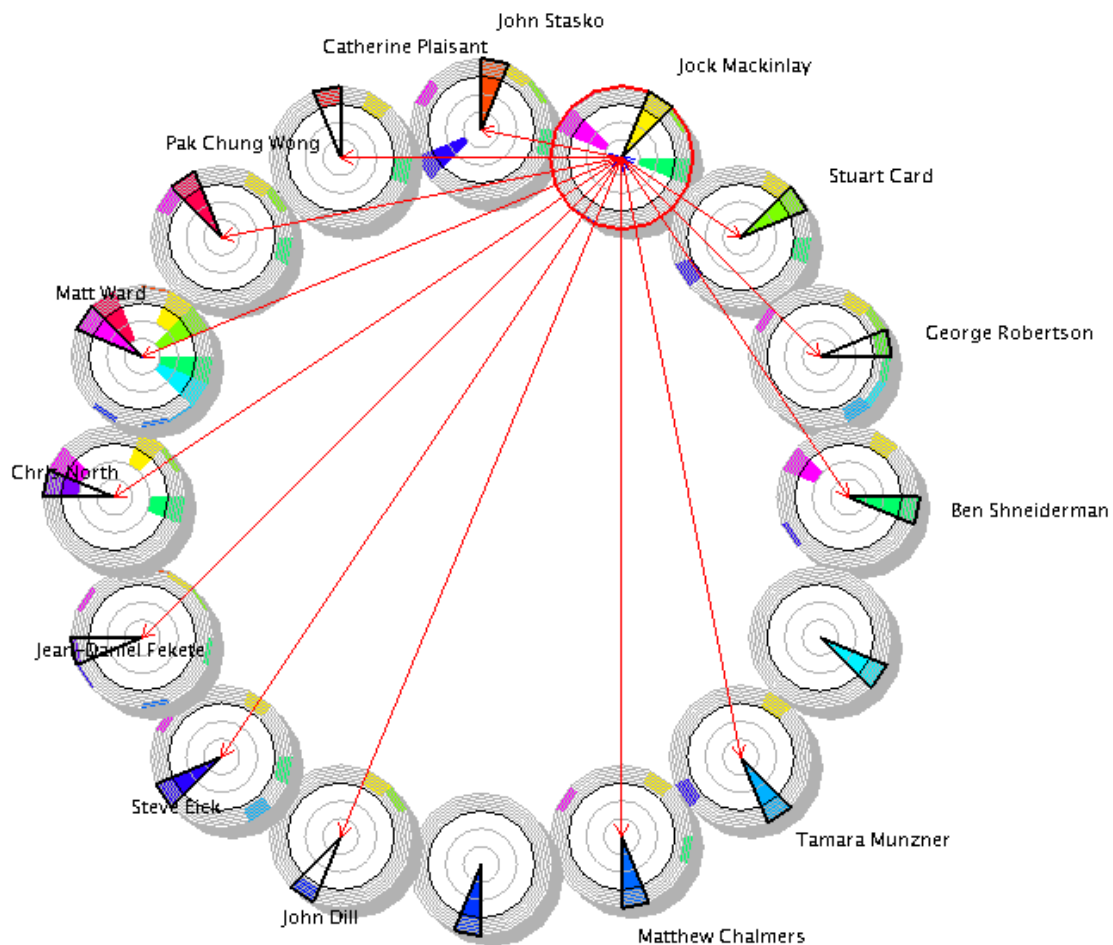


Abbildung 3.10.: Eine Influence-Visualisierung aus CiteWiz

Quelle: Elmqvist und Tsigas (2007)

Die Autoren beschreiben die dargestellten Informationen folgendermaßen: Jede der im Kreis angeordneten Scheiben repräsentiert einen Autor des Subsets. Die Scheiben sind in Anlehnung an die Jahresringe eines Baumes aufgebaut. Jeder abgetrennte Kreis innerhalb einer Scheibe repräsentiert einen Zeitabschnitt in Monaten oder Jahren, wobei sich im Inneren „jüngere“ Daten als im Äußeren befinden. Die äußeren Zeitabschnitte sind zur besseren Übersichtlichkeit komprimiert dargestellt, sodass der Fokus auf den ältesten drei Zeitscheiben im Inneren liegt. Außerdem besitzen die Scheiben Eigenschaften von

Kreisdiagrammen. Jedes Kreissegment bezieht sich auf Publikationen des Autors mit der entsprechenden Farbe, wobei jedem Autor eine individuelle Farbe (schwarz umrandet) zugeordnet ist (Elmqvist und Tsigas, 2007).

Aus der Darstellung lassen sich also detaillierte Informationen über Autoren und deren gegenseitige Zitationen gewinnen, wobei die Daten durch einen zeitlichen Wandel strukturiert werden. In der Abbildung 3.10 wurde der Autor *Jock Mackinlay* hervorgehoben, sodass alle Autoren die Diesen zitiert haben mit roten Pfeilen gekennzeichnet wurden. Dem Autor *Jock Mackinlay* selbst ist der Beschreibung zufolge die Farbe gelb zugeordnet. Das bedeutet: Alle Autoren die *Jock Mackinlay* zitiert haben, besitzen zu den Veröffentlichungszeitpunkten gelbe Anteile in ihrem Kreisdiagramm. So ist zum Beispiel zu sehen, dass nahezu alle Autoren in irgendeiner ihrer Publikationen *Jock Mackinlay* zitiert haben. Umgekehrt hat Mackinlay häufig die Autoren *Matt Ward* (pink) und *Ben Shneiderman* (grün) zitiert. Es lassen sich so mit dieser Influence-Visualisierung Details über die Relevanz eines Autors ableiten. Wird dieser über längere Zeiträume häufig zitiert, so liegt die Vermutung nahe, dass seine Veröffentlichungen von besonderer Relevanz für das Themengebiet sind. Außerdem kann durch häufige gegenseitige Zitationen eine Community von Autoren identifiziert werden.

Zusammenfassend ist zu sagen, dass die von den CiteWiz-Entwicklern erdachten Konzepte sehr robust sind und einen zusätzlichen Nutzen zu den von Ihnen genannten Basisfunktionen liefern. Das bestätigen die Darstellungen und Schlussfolgerungen im Journalartikel (Elmqvist und Tsigas, 2007), sowie die dort detailliert beschriebenen Benutzertests, ebenso wie der Gehalt an Information in den Visualisierungen. Eine umfängliche und objektive Bewertung ist leider nicht möglich, da das Programm nicht zum Testen vorliegt. Nach den gegebenen Informationen bestehen aber insbesondere noch Defizite in der Benutzbarkeit durch ungeübte Anwender (Elmqvist und Tsigas, 2007) und in der Datenbasis, die manuell generiert werden muss. Außerdem überzeugen die Visualisierungen nicht unbedingt unter ästhetischen Gesichtspunkten. Da es sich aber erstens nur um einen Prototyp zu Forschungszwecken handelt und zweitens die Entwicklung über sieben Jahre zurückliegt, sind diese Mängel vernachlässigbar. Lobenswert sind auch die recht unbekannten aber fundierten Ansätze der Visualisierung. Es liegt nahe, dass die von der Studie als positiv bewerteten Basisfunktionen in die Anwendungsfälle für AcaVis einfließen könnten. Diese Anwendungsfälle werden später in Kapitel 5 vorgestellt.

CitNetExplorer

Das Programm *CitNetExplorer* (Citation Network Explorer) wurde Anfang 2014 in der ersten Version veröffentlicht. Entwickelt wurde es durch Nees Jan van Eck und Ludo Waltman an der Universität Leiden. Die in der Programmiersprache Java geschriebene Desktopanwendung benötigt mindestens eine Java-Laufzeitumgebung der Version 6. CitNetExplorer ist kostenlos auf der offiziellen Webseite²⁹ verfügbar. Die Analyse und Visualisierung von Publikationen stehen im Mittelpunkt. Verwendet werden kann das Programm zum Beispiel, um die historische Entwicklung einer Forschungsdisziplin darzustellen oder um Publikationen von herausragender Bedeutung für ein Fachgebiet zu

²⁹<http://www.citnetexplorer.nl/Home>

identifizieren. Die eigentliche Visualisierung besteht aus einem Publikations-Netzwerk mit gegenseitigen Zitationen der Publikationen untereinander, im Verbund mit einer historischen Anordnung (van Eck und Waltman, 2014).

Es können ausschließlich bibliografische Datensätze im ISI-Format verarbeitet werden. Das Dateiformat ist gängig beim Export von bibliografischen Datensätzen aus der Web of Science Datenbank. Teilweise bieten auch andere Suchmaschinen den Export von Ergebnissen in diesem Format an. Die Daten müssen vollständig manuell generiert werden, CitNetExplorer besitzt keine direkte Anbindung an Publikationsdatenbanken. Die Datensätze werden im Programm importiert und direkt verarbeitet. Es werden Informationen über die Autoren, den Titel, das Erscheinungsdatum, gegebenenfalls das publizierende Journal und die Zitationen der Publikationen untereinander aus den bibliografischen Datensätzen extrahiert.

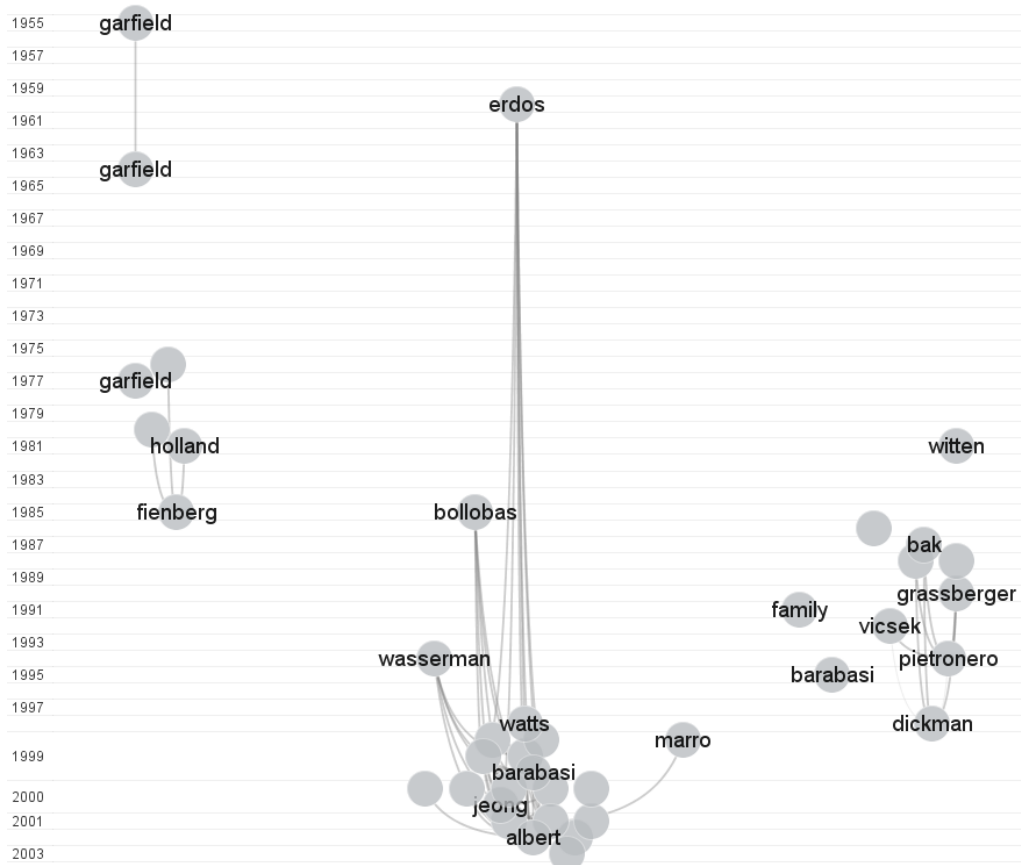


Abbildung 3.11.: Ein Netzwerk aus 40 Publikationen und deren Zitationen untereinander
 Datenquelle: Beispieldatensatz des Sci²Tool
 Visualisierung: eigenes Werk, Screenshot des Programms CitNetExplorer

Basierend auf den Zitationen wird ein Netzwerk aus Publikationen und deren Zitationen generiert. Eine beispielhafte Visualisierung eines solchen Netzwerks von 426 Publikationen

ist in der Abbildung 3.11 zu sehen. Lediglich die 40 Publikationen mit der größten Relevanz werden angezeigt, die Reduktion der Darstellung verbessert die Übersichtlichkeit. Die Relevanz ergibt sich aus einem durch Web of Science ermittelten Wert.

Jede Publikation wird durch einen Knoten im Netzwerk abgebildet. Kanten zwischen diesen Knoten stehen jeweils für die Zitationen zwischen den Publikationen. Die „Leserichtung“ der Zitationen wird durch die zeitliche Unterteilung verdeutlicht. Die Publikationen sind vertikal nach ihrem Erscheinungsdatum sortiert, wobei alte Publikationen im oberen Bildabschnitt zu sehen sind. Da nur bereits existierende Publikationen zitiert werden können, ist die „Leserichtung“ der Beziehungen also von unten nach oben.

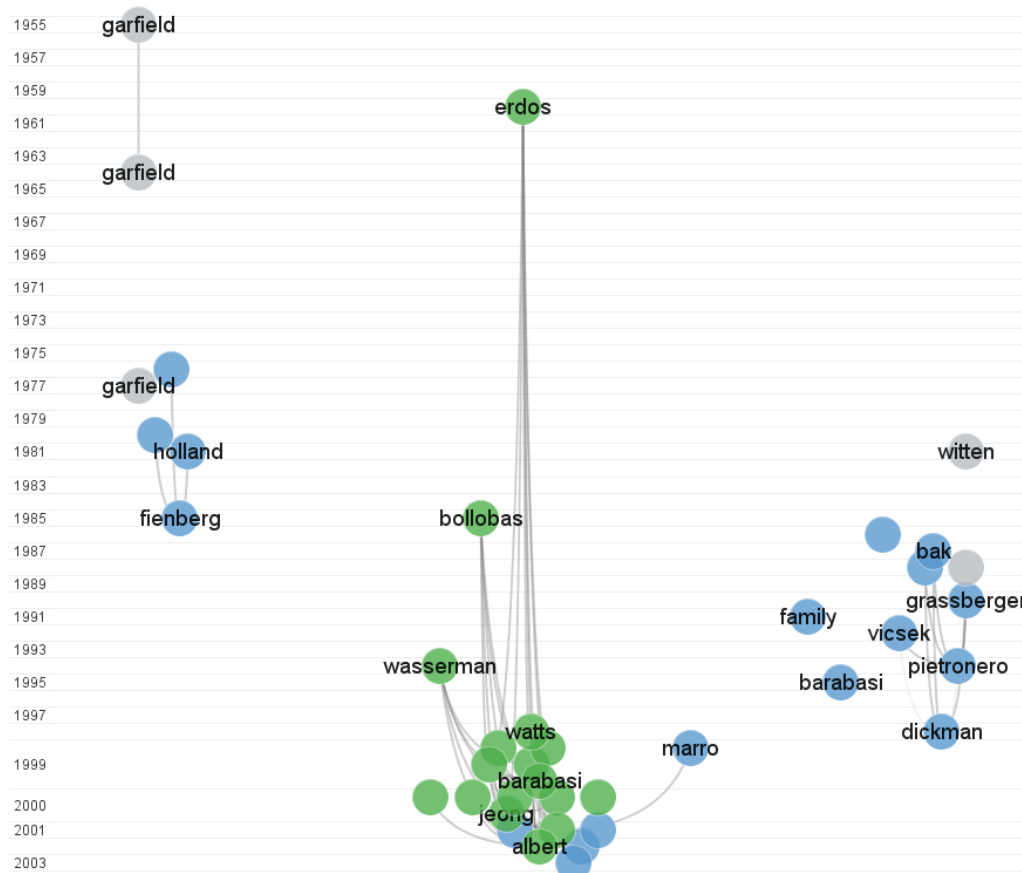


Abbildung 3.12.: Publikationen mit mindestens acht Zitationen sind blau und mit mindestens zehn Zitationen grün gefärbt
Datenquelle: Beispieldatensatz des Sci²Tool
Visualisierung: eigenes Werk, Screenshot des Programms CitNetExplorer

Ergänzend zur Visualisierung werden einige Zusatzinformationen bereitgestellt. Die Anzahl aller Publikationen und Zitationen, inklusive der nicht angezeigten Publikationen sowie die Zeitspanne der Darstellung sind verfügbar. Interaktiv werden zusätzliche Informationen über jede Publikation angezeigt, wenn diese mit dem Mauszeiger überfahren

wird. Weitere Interaktionen wie die Vergrößerung eines Teilausschnitts und Navigation in der Visualisierung sind möglich.

Ferner können die Netzwerke hinsichtlich ihrer Cluster und der Kernpublikationen sowie weiterer Eigenschaften analysiert und visualisiert werden. In der Abbildung 3.12 wurden zum Beispiel die Kernpublikationen farblich hervorgehoben. Kernpublikationen sind nach der Umsetzung von CitNetExplorer die am häufigsten zitierten Publikationen. Dabei steht eine blaue Färbung für mindestens acht Zitationen und eine grüne Färbung für mindestens zehn Zitationen dieser Publikationen. Die Anzahl der Zitationen bezieht sich auf den importierten Datensatz, nicht auf die allgemeine Anzahl der Zitationen einer Publikation. Auch ausgeblendete Knoten werden bei der Berechnung berücksichtigt, daher stimmt die Anzahl der dargestellten Zitationen als Kanten teilweise nicht mit der Färbung überein.

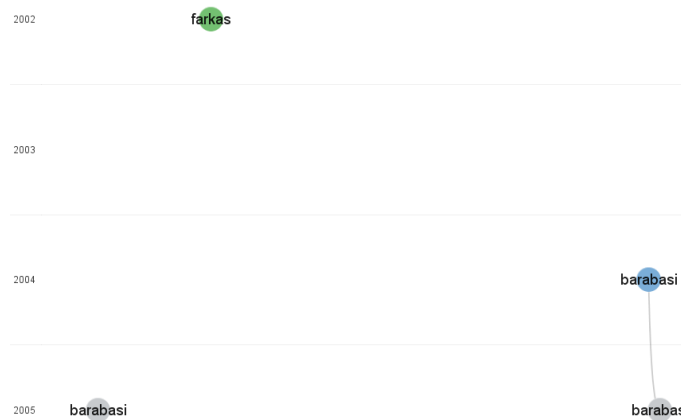


Abbildung 3.13.: Nur Publikationen mit dem Wort *network* im Titel werden dargestellt
Datenquelle: Beispieldatensatz des Sci²Tool
Visualisierung: eigenes Werk, Screenshot des Programms CitNetExplorer

Die vorgestellten Beispiele visualisieren jeweils alle Publikationen, eingeschränkt durch die genannte obere Grenze. Es ist jedoch möglich, auch nur Teilmengen dieser Publikationen zu visualisieren. Diese Teilmengen können manuell bestimmt oder durch Filterung der Publikationen gebildet werden. Gefiltert werden kann nach den Autoren, dem Titel oder dem publizierenden Journal. Auch kann der Zeitbereich eingegrenzt werden. In der Abbildung 3.13 sind beispielhaft nur Publikationen visualisiert, in deren Titel das Wort *network* auftaucht.

Im nächsten Beispiel wurden die Publikationen durch den integrierten Algorithmus in Cluster eingeteilt. Die Anzahl der dargestellten Publikationen wurde auf 60 erhöht. Insgesamt zehn Cluster wurden identifiziert. Der Algorithmus berücksichtigt die Quantität der Zitationen bei der Einteilung in Cluster.

Besonders interessant an der Visualisierung sind die verschiedenen isolierten Gruppen von Publikationen. Unter der Voraussetzung, dass die gesammelten Publikationen ein umfangreiches Gesamtbild einer Disziplin oder eines Themengebiets zeigen, lässt sich so Grundlagenliteratur anhand der Zeit und der Quantität der Zitationen identifizieren. Es

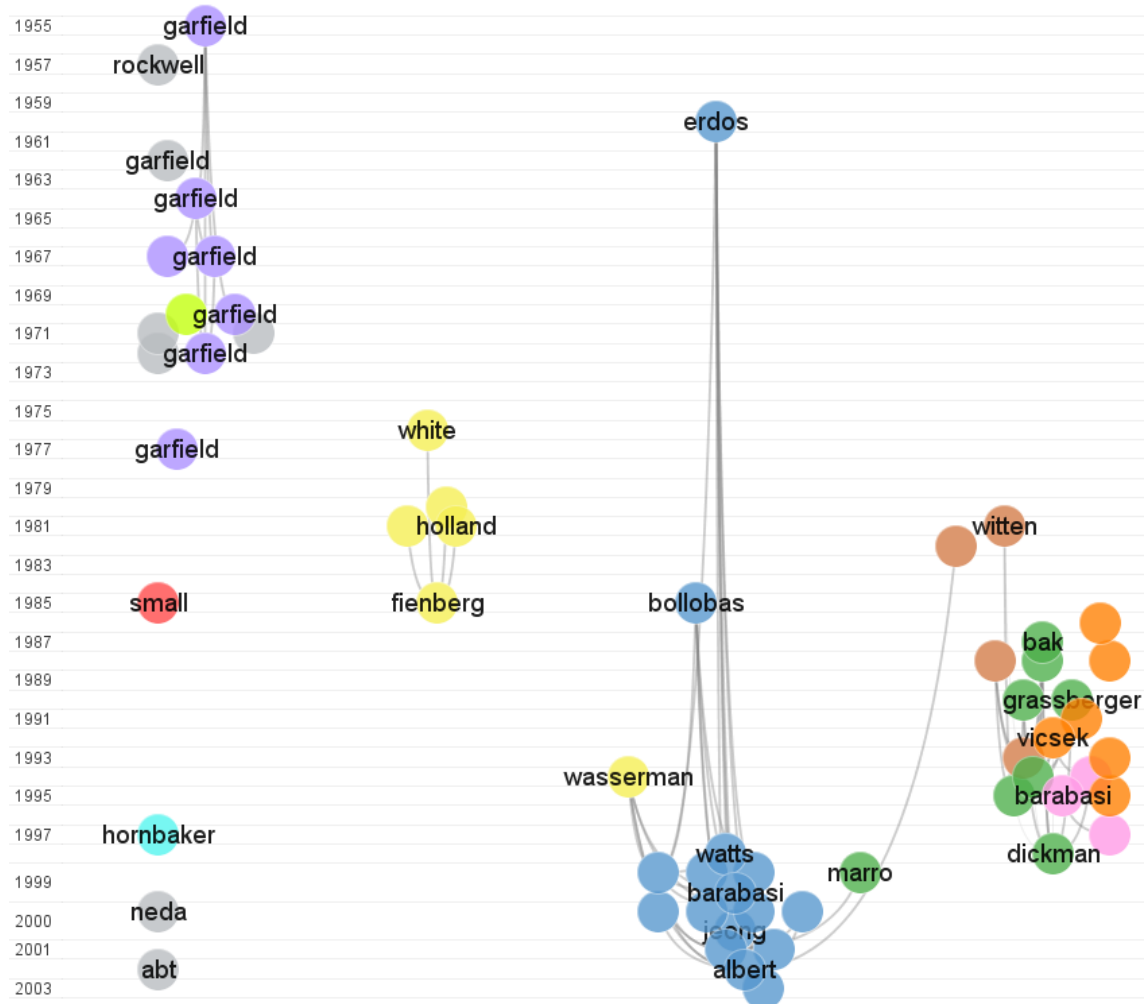


Abbildung 3.14.: 60 Publikationen eingeordnet in 10 Cluster, jede Farbe repräsentiert einen anderen Cluster
Datenquelle: Beispieldatensatz des Sci²Tool
Visualisierung: eigenes Werk, Screenshot des Programms CitNetExplorer

lässt sich durch die zeitliche Einordnung beispielsweise auch ableiten, ob ein Themengebiet noch Gegenstand aktueller Forschung ist.

Die vorgenommenen Manipulationen am Netzwerk können abgespeichert und später wieder aufgerufen werden. Auch kann die Visualisierung direkt als Raster- oder Vektorgrafik exportiert werden.

CitNetExplorer überzeugt insbesondere in seiner einfachen Bedienung und Effektivität. Die Benutzeroberfläche ist klar geordnet und intuitiv aufgebaut. Die Netzwerkvisualisierung zeichnet sich besonders durch ihre minimalistische Darstellung der wichtigsten Aspekte aus. Gelungen sind auch die Interaktionsmöglichkeiten mit der Visualisierung, da sie eine sehr intuitive Navigation ermöglichen. Die Funktionen der Filterung und Analyse

beschränken sich dabei auf die wesentlichen Aspekte. Die Visualisierung steht immer im Mittelpunkt und wird durch diese Funktionen punktuell verfeinert. Viele kleinere Hilfsfunktionen können mit der Zeit entdeckt werden und erlauben dadurch eine Fortentwicklung in der Art und Weise der Benutzung. Die Arbeit mit dem Programm umfasst nur einen wesentlichen Ablauf, sodass ein spielerisches Experimentieren und eine Exploration der Funktionen möglich sind. Auch Personen die keine besonderen Kenntnisse über die Theorie der bibliografischen Organisation und Visualisierung besitzen, sind in der Lage, in kurzer Zeit aussagekräftige Ergebnisse zu erzielen. Die Unterstützung einer größeren Anzahl von Dateiformaten wäre allerdings wünschenswert. Insgesamt hinterlässt CitNet-Explorer einen sehr positiven Eindruck. Es wird exakt ein Lösungsansatz angeboten und dieser sehr gut umgesetzt. Dies ist bemerkenswert, da es sich um die erste Version einer vollständigen Neuentwicklung handelt. Ein ähnlich gearteter Typ von Visualisierung könnte unter Umständen auch innerhalb von AcaVis zum Einsatz kommen.

VOSviewer

Das Programm VOSviewer wurde im Jahr 2009 von Nees Jan van Eck und Ludo Waltman vorgestellt. Derzeit ist die Version 1.5.7 aktuell, welche kostenlos auf der Webseite³⁰ heruntergeladen werden kann. Das Programm ist in der Programmiersprache Java geschrieben und benötigt mindestens eine Java-Laufzeitumgebung der Version 6. VOSviewer wurde entwickelt, um bibliometrische Karten und Graphen darzustellen (van Eck und Waltman, 2010). Benutzungsszenarien umfassen nach van Eck und Waltman (2010) zum Beispiel die Darstellung von Autoren und Journalen basierend auf Co-Zitationen oder die Analyse von Volltexten nach Schlagworten, die gemeinsam auftreten (Co-Occurrence). Auf Basis der Daten können zum Beispiel Graphen und Heatmaps erzeugt werden, die bestimmte Strukturen offenbaren. Das Programm dient einerseits dem Zweck, bibliografische Daten und Schlagworte für eine Analyse zu visualisieren und andererseits dem Zweck einer ersten praktischen Umsetzung des VOS-Algorithmus, wie er von van Eck und Waltman (2007) entwickelt wurde. Um einen Einblick in den Arbeitsablauf mit VOSviewer zu geben, werden im Nachfolgenden die Kernfunktionen des Programms erläutert. Auch werden Beispieldaten und deren Visualisierungen mit VOSviewer zur Verdeutlichung verwendet.

Für die Analyse von Co-Zitationen können Daten in den Exportformaten von Web Of Science und Scopus direkt verwendet werden. Außerdem können die Daten aus für verschiedene Zwecke gebräuchlichen Dateiformaten des Programms Pajek importiert werden. Für die Schlagwortanalyse können Volltexte aus Textdateien importiert werden. Zusätzlich ist es bei allen Importmethoden möglich, Textdateien mit Thesaurus-Begriffen für die Wortnormalisierung anzugeben. VOSviewer besitzt keine Online-Schnittstelle zu den genannten oder anderen Datenbanken, die Datenbasis muss also manuell erstellt werden. Nach dem Import werden die Daten automatisch durch den verwendeten VOS-Algorithmus verarbeitet und anschließend dargestellt. Die Anordnung der Autoren, Titel oder Schlagworte als Knoten erfolgt durch eine Distanz-Darstellung. Je häufiger Co-Zitationen oder Co-Occurrence der Einheiten sind, desto näher zusammen werden diese dargestellt. Außerdem ist es möglich, anhand der Werte eine Einteilung der Einheiten in Cluster berechnen zu lassen.

³⁰VOSviewer Webseite <http://www.vosviewer.com/>

Als Datenbasis für die folgenden praktischen Beispiele wird ein Datensatz der 1.242 meist zitierten Publikationen aus dem Bereich der Informationswissenschaften der Jahre 1999 bis 2008 verwendet. Dieser ist neben weiteren Datensätzen online auf der Webseite von VOSviewer erhältlich³¹. Der Datensatz wurde auch im Journalartikel *A unified approach to mapping and clustering of bibliometric networks* von Waltman et al. (2010) verwendet und im Detail besprochen. Das methodische Vorgehen der Datensammlung umfasst einige weitere Aspekte, insbesondere die Eingrenzung auf bestimmte Publikationen, die in dem Journalartikel in Kapitel 4 nachzulesen sind. Die Kenntnis ist für die nachfolgenden Betrachtungen nicht notwendig.

In der Abbildung 3.15 wurde der beschriebene Datensatz visualisiert. Jeder Knoten steht für eine Publikation, wobei die Publikationen mit dem größten Gewicht, basierend auf den Co-Zitationen, mit dem entsprechenden Erstautor beschriftet sind. Die Einteilung der insgesamt 25 Cluster wurde mit den von Waltman et al. (2010) verwendeten Parametern durchgeführt. Für jeden Cluster werden die Knoten in einem anderen Farbton dargestellt.

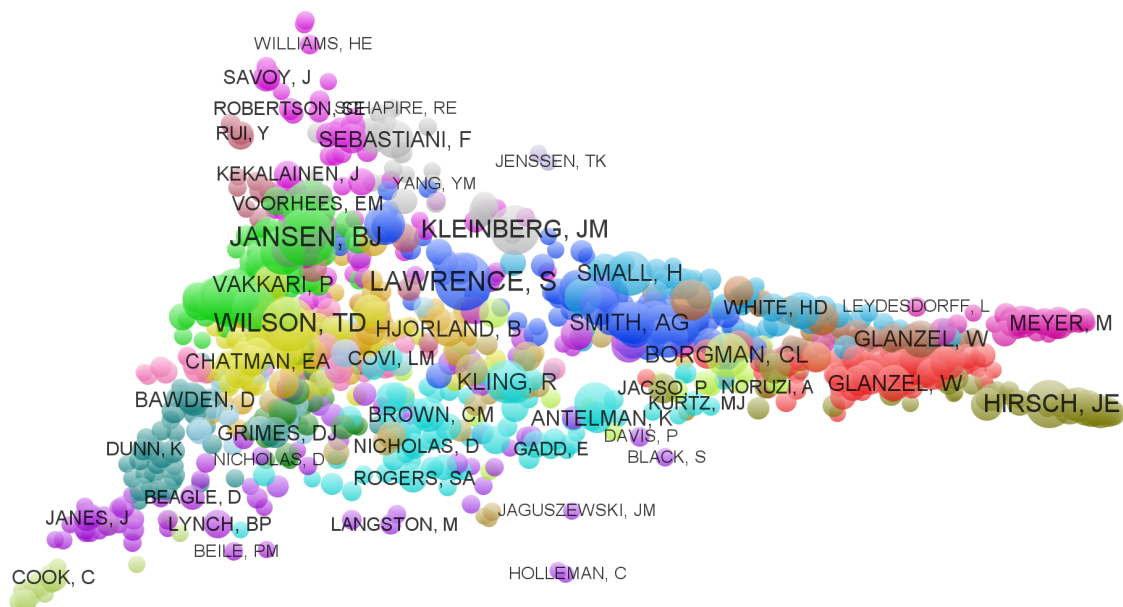


Abbildung 3.15.: Karte von 1.242 Publikationen, eingeteilt in Cluster, basierend auf Co-Zitationen. Größere Knoten stellen besonders häufig zitierte Publikationen dar.

Datenquelle: Waltman et al. (2010)

Visualisierung: eigenes Werk, Screenshot des Programms VOSviewer

Diese Darstellung ermöglicht es dem Betrachter, sich zunächst einen groben Überblick über das verwendete Zitations-Netzwerk zu verschaffen. Mehrere Formen der Interaktion mit der dargestellten Karte sind möglich. Allgemein ist das Vergrößern von beliebigen Kartenausschnitten möglich. Der Bildausschnitt kann entsprechend verschoben werden.

³¹Beispieldatensatz http://www.ludowaltman.nl/unified_approach/

Es ist ebenfalls möglich, Detailinformationen zu einer Publikation anzuzeigen, wenn diese mit dem Mauszeiger überfahren wird. Außerdem lässt sich die Anzahl der angezeigten Knoten individuell anpassen. Weitere Optionen zur Modifikation der optischen Attribute der Visualisierung stehen zur Verfügung.

Mit näherer Kenntnis über das Fachgebiet beziehungsweise über die Publikationen lassen sich einige Aussagen aus der Visualisierung gewinnen. Der linke Teil der Karte umfasst viele Publikationen, die zum Teilbereich *Information Retrieval* gezählt werden und Publikationen im rechten Teil werden eher dem Teilbereich *Infometrics* zugeordnet, wie Waltman et al. (2010) beschreiben.

In der folgenden Abbildung 3.16 ist der gleiche Datensatz visualisiert, jedoch als Heatmap. Bereiche mit besonders vielen Publikationen werden in rot dargestellt, Bereiche mit wenigen Publikationen in grün und blau. Die entsprechenden Übergänge sind manuell anzupassen. Bereiche mit vielen Publikationen kommen dadurch zustande, dass diese allgemein häufig zitiert werden und sich in einigen Fällen oft untereinander zitieren. Zur Orientierung wurden wieder die Erstautoren der wichtigsten Publikationen eingefügt, wobei in dieser Abbildung mehr Beschriftungen enthalten sind, als in der Abbildung 3.15. In dieser Darstellung der gleichen Publikationen lassen sich noch besser als zuvor die beiden genannten Teilgebiete identifizieren. Entsprechend der Heatmap-Metapher sind dies die „heißen“ Bereiche.

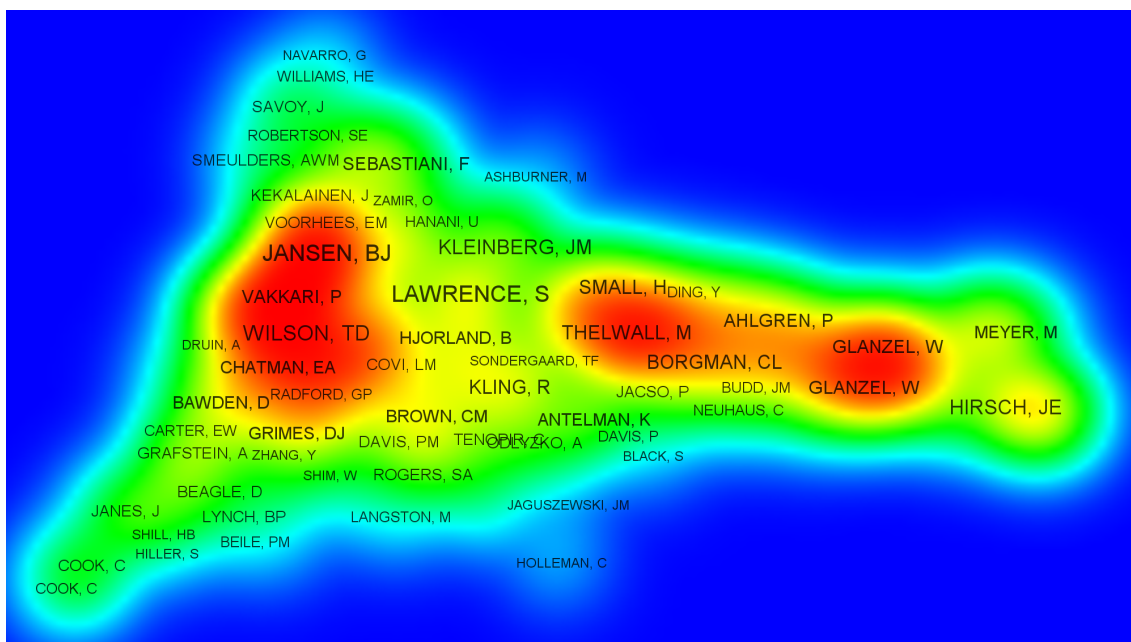


Abbildung 3.16.: Heatmap von 1.242 Publikationen. Bereiche mit vielen Publikationen sind in rot dargestellt und Bereiche mit wenigen Publikationen sind grün und blau dargestellt.

Datenquelle: Waltman et al. (2010)

Visualisierung: eigenes Werk, Screenshot des Programms VOSviewer

In der dritten Darstellung lassen sich sämtliche Publikationen im Detail erkennen. Sie sind jeweils durch einen Punkt in der Abbildung 3.17 dargestellt. Die Farbe jedes Punkts in dieser Punktwolke ordnet diesen zu einem der 25 Cluster mit dem jeweiligen Farbton zu. Diese Form der Darstellung ist im Gegensatz zu den beiden vorherigen Darstellungen schlecht dazu geeignet, Aussagen über das gesamte Netzwerk zu treffen. Vielmehr kann hier jede einzelne Publikation für sich betrachtet werden, da sich die Punkte nicht gegenseitig überdecken. Durch Benutzerinteraktion mit dem Mauszeiger und der Darstellung können ebenfalls erweiterte Informationen über eine Publikation angezeigt werden. Um den Detailcharakter der Darstellung zu verdeutlichen, zeigt die Abbildung 3.17 auch nur den oberen linken Ausschnitt des gesamten Netzwerks in vergrößerter Ansicht.

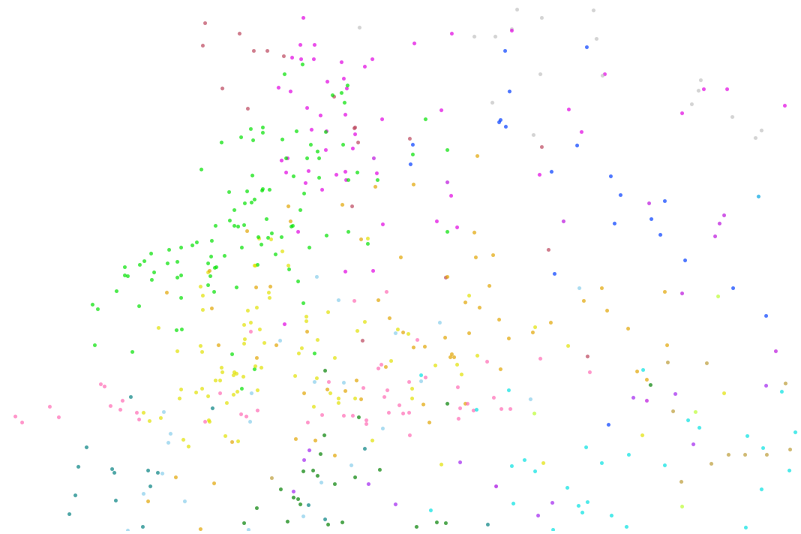


Abbildung 3.17.: Punktwolke von 1.242 Publikationen. Jeder Punkt stellt eine Publikation dar und ist in der Farbe des jeweiligen Clusters dargestellt.

Datenquelle: Waltman et al. (2010)

Visualisierung: eigenes Werk, Screenshot des Programms VOSviewer

Neben den genannten Optionen bestehen noch viele kleinere Einstellmöglichkeiten. Die Meisten sind nicht elementar und erlauben es Experten, bestimmte Details in der Visualisierung herauszustellen. Die erstellten Netzwerke und Cluster lassen sich mit VOSviewer auch exportieren, sodass die Ergebnisse später erneut betrachtet oder verändert werden können. Auch der Export der einzelnen Visualisierungen als Raster- oder Vektorgrafik ist möglich.

VOSviewer stellt eine durchdachte, abgerundete Lösung zur Visualisierung großer bibliografischer Datensätze bereit. Die enthaltenen Methoden sind für diesen Einsatz gut gewählt, weil besonders geeignet. Der Arbeitsablauf ist klar und bedarf wenig Vorwissen. Ein spielerisches Einarbeiten ist möglich, ansprechende Ergebnisse können in kurzer Zeit erstellt werden. Insbesondere sind die verwendeten Verfahren immer für minimale Beispiele geeignet, da die Parameter sinnvoll mit Standardwerten belegt sind. Erfahrenen Benutzern ist es durch die große Anzahl an Optionen möglich, bedeutende Verfeinerungen

vorzunehmen. Das Programm erreicht die anspruchsvollen Ziele, die durch die Entwickler von VOSviewer gesetzt wurden. Es wäre hier aber ebenfalls eine direkte Anbindung an bibliografische Online-Datenbanken wünschenswert.

3.2. Ansätze zur Recherche und Organisation

Die vorgestellten Ansätze der Visualisierung sind zwar gut geeignet, um bibliografische Daten und Metriken darzustellen, lassen jedoch allesamt einen Aspekt vermissen: Die direkte Anbindung an eine Online-Datenbank für nicht-fachgebundene bibliografische Daten. Deshalb soll im folgenden Abschnitt der umgekehrte Fall untersucht werden, ob die verfügbaren bibliografischen Online-Datenbanken auch Möglichkeiten zur Visualisierung der Daten bereitstellen. Dazu wird jeweils eine kurze Beschreibung einiger Online-Datenbanken gegeben und die dort verfügbaren Visualisierungen werden evaluiert. Neben den gezeigten Vertretern gibt es noch weitere Datenbanken. Die Vorliegenden wurden wegen ihres Bekanntheitsgrads oder wegen der umgesetzten Visualisierungen ausgewählt.

Vorab ist anzumerken, dass das Mendeley-Projekt eine Sonderrolle einnimmt, da es sich dabei in den Ursprüngen um ein desktopbasiertes System zur bibliografischen Organisation handelte. Hinzugekommen ist jedoch der Mendeley Online-Katalog für Publikationen. Bei den restlichen Systemen handelt es sich vorrangig um Suchmaschinen für Publikationen, bei denen die bibliografische Organisation nur beiläufig oder gar nicht zum Funktionsumfang gehört.

3.2.1. Mendeley

Mendeley ist eine Software zur Verwaltung von bibliografischen Sammlungen und deren Metadaten. Verknüpft mit einem Benutzerzugang zu Mendeley können die eigenen Sammlungen web- und desktopbasiert abgerufen werden. Die Registrierung eines Benutzerzugangs ist kostenlos, für größere Datenmengen kann der Zugang gegen eine Gebühr aufgewertet werden. Als umfängliches bibliografisches Organisationssystem ist Mendeley in der Lage, Volltexte und Metadaten zusammenzuführen. Die üblichen öffentlich zur Verfügung stehenden Informationen können durch Anmerkungen und Schlagworte individuell erweitert werden, sodass der Fokus auf den eigenen Forschungsinteressen liegt. Neben der Funktion als Verwaltungssystem hat sich Mendeley zu einer Art sozialen Plattform für Forschende entwickelt. Die Möglichkeit sich in Interessengruppen einzutragen und von diesen Gruppen Empfehlungen für lesenswerte Publikationen zu erhalten, stärkt den kollaborativen Charakter.

Genauer betrachtet werden soll nun die über Mendeley verfügbare Datenbasis. Die integrierte Suchfunktion kann zum Auffinden von Publikationen und Autoren verwendet werden. Dabei werden nach eigenen Angaben³² die Metadaten und gegebenenfalls vorhandene Volltexte durchsucht und nach einem nicht näher beschriebenen Ranking-Algorithmus ausgegeben. Neben dieser Textsuche gibt es die Möglichkeit, Publikationen anhand von Schlagworten zu finden³³. Die Schlagworte werden von Nutzern den Publikationen zuge-

³²<http://blog.mendeley.com/press-release/mendeley-catalog-search-engine/>

³³<http://blog.mendeley.com/progress-update/mendeleys-research-catalog-is-now-wikified-come-help-us-organize-the-worlds-research/>

ordnet, sodass deren Integrität in jedem Einzelfall kritisch betrachtet werden sollte. Die öffentlichen Schlagworte sind unabhängig von privat verfassten Anmerkungen und Schlagworten³³.

Eine klassische Visualisierung von Publikations-Metadaten seitens Mendeley ist nicht vorhanden. Die einzige visuelle Darstellung ist eine Übersicht über den Zugriff auf eine Publikation bezüglich des Herkunftslands, des akademischen Grads und der Forschungsdisziplin der Leser einer Publikation. Es existieren allerdings zwei experimentelle Ansätze der Visualisierung mit Unterstützung der Mendeley-Forschungsabteilung *Mendeley Labs*.

Der erste Ansatz ist das Projekt *Mendeley Cross-Country Collaborations*, welches eine Darstellung der internationalen Zusammenarbeit von Mendeley-Benutzern beinhaltet³⁴. Die zugrundeliegenden Daten sind lediglich eine Momentaufnahme aus dem Jahr 2013. Die Darstellung³⁵ beinhaltet eine interaktive Weltkarte, die durch Verbindungen zwischen den Ländern deren Zusammenarbeit zeigt. Eine Aufnahme der Karte ist in der Abbildung 3.18 zu sehen. Außerdem ist eine tabellarische Darstellung der am häufigsten verlinkten Länder in der Visualisierung enthalten. Kritisch betrachtet werden müssen die Bedingungen, unter denen die Daten erhoben wurden³⁴. Außerdem ist die Aussagekraft der Daten für die reale internationale Zusammenarbeit von Forschern zu hinterfragen. Erstens stellt die Mendeley-Community nicht unbedingt einen repräsentativen Querschnitt von Forschern dar und zweitens impliziert die Vernetzung von Forschern über das Internet keine reale Vernetzung und vice versa. Trotz der Defizite ist dies immerhin ein Ansatz, der für die Visualisierung aufgegriffen werden kann. Diese Form der Visualisierung genießt offenbar ein gewisses Publikum³⁴.

Das zweite Projekt mit dem Namen *Head Start* versucht Schlüsselpublikationen für wissenschaftliche Disziplinen zu visualisieren. Als Metrik für die Relevanz einer Publikation dienen die gesammelten Daten über die Leser der Publikation in Mendeley. Das methodische Vorgehen wird von Kraker et al. (2013) in ihrem Paper *Head Start: Improving Academic Literature Search with Overview Visualizations based on Readership Statistics* beschrieben. Aus dem Datenbestand von Mendeley wurden 91 relevante Publikationen der wissenschaftlichen Disziplin *Educational Technology* ausgewählt. Die Daten zeigen ebenfalls eine Momentaufnahme des Jahres 2013. Die Publikationen wurden von den Autoren des Papers anhand ihrer Titel analysiert und entsprechenden Disziplinen zugeordnet. Die globale Einordnung der 91 Publikationen ist in der Abbildung 3.19 zu sehen.

Die Darstellung lässt sich interaktiv auf die einzelnen Teilgebiete reduzieren. Die angezeigten Publikationen können durch die Eingabemaske auf der rechten Seite gefiltert werden. Ein Klick auf eine Publikation öffnet diese in der Weboberfläche von Mendeley. Die Visualisierung ist online über den Server von Mendeley Labs verfügbar³⁶. Die Motivation dieser Visualisierung ist das wiederkehrende Problem bei der wissenschaftlichen Recherche, die Relevanz einer Publikation einzuordnen (Kraker et al., 2013). Eine Erstpublikation ist immer die exakteste Quelle, im Bezug auf die dortigen Ergebnisse. Zusammenfassungen und aufbauende Arbeiten geben die Inhalte nicht vollständig oder

³⁴<http://blog.mendeley.com/tag/cross-country-collaboration/>

³⁵Cross-Country Collaboration Map <http://labs.mendeley.com/collab-map/>

³⁶Head Start Projekt <http://labs.mendeley.com/headstart/>

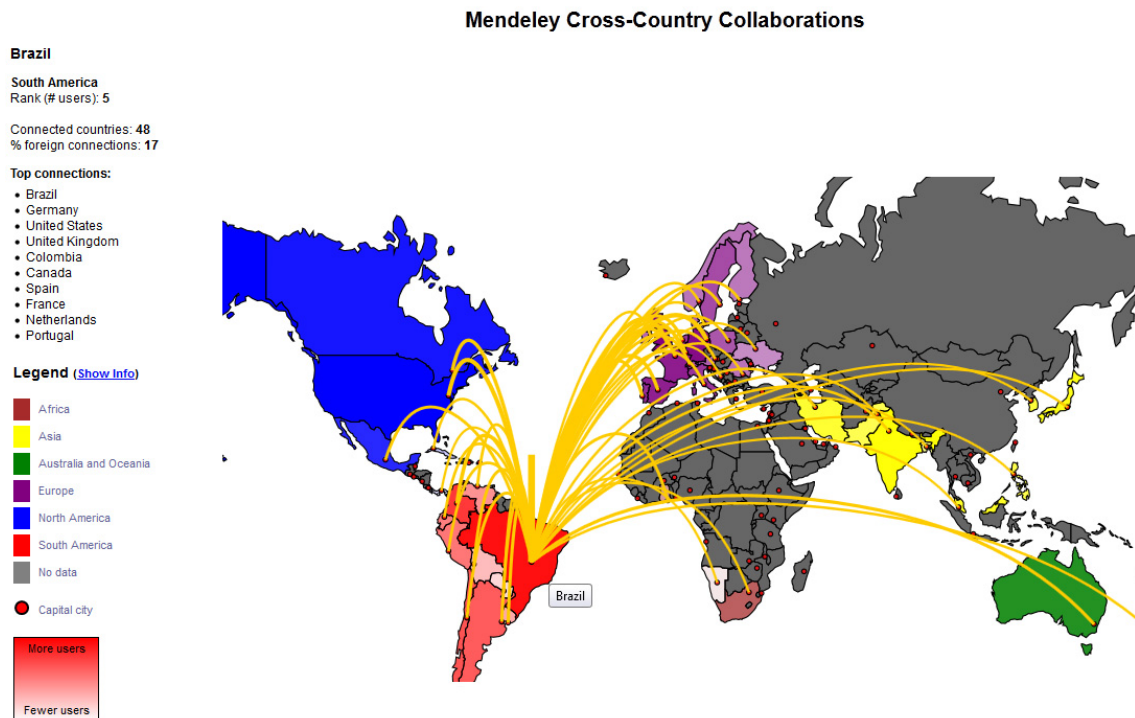


Abbildung 3.18.: Weltkarte der internationalen Zusammenarbeit in Mendeley. Zu sehen sind die internationalen Verbindungen von Brasilien.
Quelle: Mendeley Labs³⁵, Screenshot der Webseite

schlimmstenfalls fehlerhaft wieder. Als Metrik die Anzahl der Leser in Mendeley heranzuziehen ist meiner Meinung nach gerechtfertigt, da davon ausgegangen werden kann, dass innerhalb einer Forschungsdisziplin der Querschnitt von Forschern zumindest näherungsweise repräsentativ ist. Untermuert wird die Aussage auch durch die Ergebnisse der eigenen Projektevaluation durch Kraker et al. (2013). Demnach seien die Ergebnisse kohärent mit den Ergebnissen anderer Studien, die für eine Identifikation der Schlüsselpublikationen im Bereich Educational Technology Zitations-Analysen verwendet haben. Die vorliegende Aufarbeitung hat ein großes Potenzial, da sie eine der wichtigen Fragen bezüglich bibliografischer Daten beantworten möchte und dies mit einem alternativen Ansatz zur Zitations-Analyse vollbringt.

Abschließend folgt eine Betrachtung der API und dem Potenzial von Mendeley. Für registrierte Benutzer stellt Mendeley eine API zur Verfügung. Diese erlaubt es unter anderem, die eigene Sammlung zu verwalten und zu betrachten, aber auch detaillierte Informationen über beliebige Publikationen im Mendeley-Katalog abzufragen³⁷. Die Datenbasis wird ausschließlich durch manuelle Einträge von Mendeley-Benutzern erstellt bzw. generiert³⁸, genauer aus deren lokalen bibliografischen Sammlungen. Durch Aggregation der

³⁷API für Publikationsmetadaten <http://dev.mendeley.com/slate/#retrieving-documents16>

³⁸Datensammlung des Mendeley-Katalogs <http://blog.mendeley.com/tag/mendeley-web-catalog/>

3. Praktische Ansätze der bibliografischen Organisation und Visualisierung

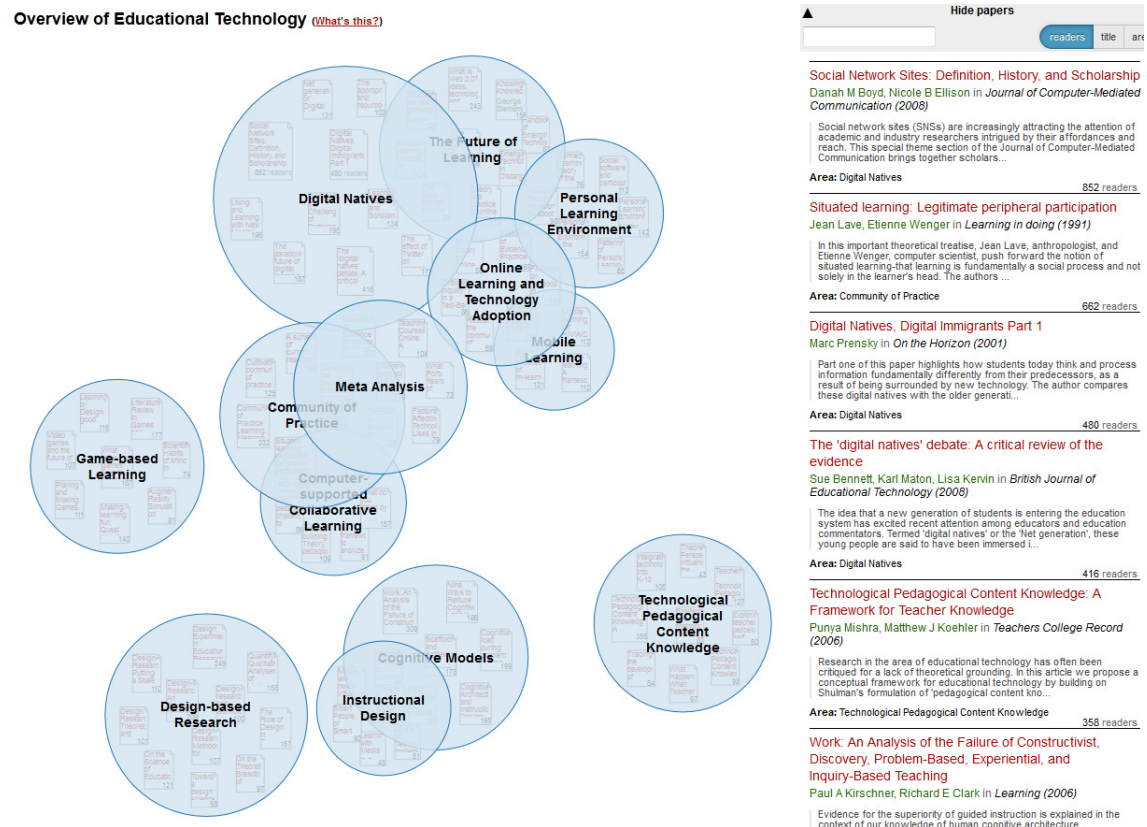


Abbildung 3.19.: Darstellung wichtiger Publikationen von Teilgebieten der Disziplin Educational Technology, aufbauend auf Daten von Mendeley
Quelle: Mendeley Labs³⁶, Screenshot der Webseite

Informationen aus den Sammlungen einer Vielzahl von Benutzern entsteht den Entwicklern zufolge eine gewisse Integrität³⁸. Es besteht die Möglichkeit, den Mendeley-Katalog nach Titeln, Autoren und Erscheinungsjahr zu durchsuchen³⁹. Mit Hilfe der API können auch Informationen über die Anzahl der Mendeley-Benutzer, die eine Publikation gelesen oder in der eigenen Sammlung gespeichert haben, abgerufen werden. Das eröffnet die Möglichkeit, für eine beliebige Menge von Publikationen aus dem Mendeley-Katalog eine weiterführende Analyse nach dem Vorgehen von Head Start (Kraker et al., 2013) durchzuführen. Dieser Ansatz wird im Abschnitt 4.1 noch einmal aufgegriffen.

Die Möglichkeiten und die API selbst sind natürlich auch für AcaVis und ähnliche Projekte relevant. Unklar ist die technische und zeitliche Umsetzbarkeit im Rahmen dieser Arbeit. Abschließend lässt sich zur Mendeley-Plattform sagen, dass das Potenzial für Visualisierungen als Hilfsmittel bei der Recherche von wissenschaftlichen Publikationen vollkommen ungenutzt bleibt. Die gezeigten experimentellen Ansätze sind durchaus vielversprechend. Es ist wünschenswert, dass derartige Erweiterungen in den Kern von Mendeley einfließen.

³⁹API für Katalogsuche <http://apidocs.mendeley.com/home/public-resources/search-terms>

3.2.2. Google Scholar

Google Scholar ist eine Auskopplung der Suchmaschine Google, mit deren Hilfe nach wissenschaftlichen Publikationen gesucht werden kann. Daneben sind seit einiger Zeit auch Patente und Gerichtsentscheidungen über Google Scholar auffindbar⁴⁰. Neben der Suchmaschine gibt es eine weitere Auskopplung namens Google Scholar Citations. Das Portal ermöglicht es Forschern, eine Übersicht über die auf ihre Arbeit referenzierenden Publikationen zu erhalten und Metriken zum eigenen Einfluss zu generieren. Es lassen sich auch öffentliche Profile im Stile eines sozialen Netzwerks erstellen, die diese Informationen allen Benutzern bereitstellen.

Google Scholar bedient sich nach eigenen Angaben⁴⁰ aus dem öffentlichen Internet durch die regulären Google-Webcrawler sowie direkt aus den internen oder bezahlpflichtigen Angeboten der Herausgeber von wissenschaftlicher Literatur. Dabei wird fast das gesamte Spektrum verschiedener Arten von Publikationen abgedeckt. Aktualisiert werden die Daten nach Angaben von Google Scholar⁴⁰ regelmäßig in Abständen von mehreren Monaten bis zu einem Jahr. Durchsuchbar sind Titel, Abstracts und Autoren der Publikationen. Die Suchfunktionen und Filter sind ähnlich umfangreich wie die der normalen Google-Suchmaschine. Google Scholar veröffentlicht keine Statistiken über die Größe des Datenbestands. Orduña-Malea et al. (2014) schätzen den Umfang basierend auf mehreren weiteren Studien auf etwa 99,3 Millionen Publikationen.

Angezeigt werden in Google Scholar, falls vorhanden, Daten über die Autoren, den Titel, Abstracts, referenzierende Publikationen und Links zu Volltexten, falls öffentlich verfügbar. Eine eigene Übersichtsseite für die jeweiligen Publikationen ist nicht verfügbar. Die Detailansicht bezieht sich auf die Webseiten der Herausgeber oder auf die Webseiten von denen aus die Publikationen indiziert wurden.

Es ist für die Publikationen keine Form der Visualisierung vorhanden. Lediglich Google Scholar Citations stellt für jede seiner Autorensseiten eine histografische Darstellung über die Zahl der Zitationen pro Jahr bereit. In der Abbildung 3.20 ist ein solches Histogramm zu sehen. Daneben sind einige Impact-Metriken sowie die Gesamtzahl der Zitationen von Publikationen des Autors gegeben.

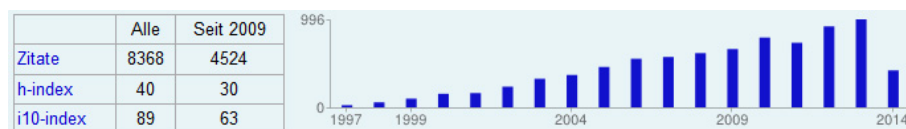


Abbildung 3.20.: Histografische Darstellung der Zitationen und Impact-Metriken eines Autors in Google Scholar Citations
Quelle: Google Scholar Citations, Screenshot

Die Suchmaschine Google Scholar wird in mehreren Papern wegen der verwendeten Methoden kritisiert. Beel und Gipp (2009) stellen in ihrem Paper *Google Scholar's ranking algorithm: The impact of citation counts* dar, welche Faktoren den verwendeten Ranking-Algorithmus stark beeinflussen. Auf Basis empirischer Studien wird darin festgestellt, dass

⁴⁰Inhalte von Google Scholar <http://scholar.google.com/intl/en/scholar/help.html#coverage>

insbesondere die Anzahl an Zitationen einer Publikation und Wortanalysen des Abstracts einen Einfluss haben. Das Ergebnis ist eine teilweise Verzerrung der Relevanz von Publikationen, wie die Autoren resümieren. In einem späteren Artikel beleuchten die Autoren die Anfälligkeit von Google Scholar für Manipulationen durch Spam. So kommen Beel und Gipp (2010) im Artikel *Academic Search Engine Spam and Google Scholar's Resilience Against it* zu dem Ergebnis, dass Google Scholar kaum Maßnahmen gegen Spam ergreift und sich damit auch die Anzahl an Referenzen auf eine Publikation manipulieren lässt. Fatal ist dies, wie die Autoren mahnen, weil erstens der Ranking-Algorithmus von Google Scholar darauf aufbaut und zweitens die Anzahl an Zitationen als gängiges Maß für den Einfluss eines Autors verwendet wird. Zusammen mit den Ergebnissen von Orduña-Malea et al. (2014), die darstellen, dass Google Scholar die vorrangig genutzte Suchmaschine für wissenschaftliche Publikationen ist, erhält die Kritik an Google Scholar weiteres Gewicht.

Die Beliebtheit von Google Scholar lässt sich insbesondere durch das vertraute, aufgeräumte Erscheinungsbild im Stile der Google-Dienste erklären. Google schlägt darüber hinaus innerhalb der normalen Suchmaschine wissenschaftliche Publikationen via Google Scholar vor, wenn diese zum Suchbegriff passen. Durch den verwendeten Ranking-Algorithmus entsteht außerdem das subjektive Gefühl, alle gezeigten Ergebnisse seien relevant und neu, was objektiv nicht immer korrekt sein muss. Beel und Gipp (2009) beschreiben dieses Phänomen ausführlich. Trotz seines Gewichts in der Branche und den damit verbundenen Möglichkeiten nutzt Google Scholar nicht das Potenzial der eigenen Daten. In den vorangegangenen Abschnitten dieser Arbeit wurden diverse Techniken gezeigt, die eine wissenschaftliche Recherche durch Visualisierungen verbessern oder beschleunigen können. Die übliche Verschwiegenheit von Google bringt es mit sich, dass auch Google Scholar kaum substanzielle Informationen über die eigenen Tätigkeiten bereitstellt. Im Fall Google Scholar kann dies auch an Vereinbarungen mit den Herausgebern bezahlpflichtiger Inhalte liegen. Die Vermutung, dass Google Scholar keine API bereitstellt, weil die nötigen Rechte zur Veröffentlichung nicht vorliegen, steht auch deshalb im Raum.

3.2.3. Microsoft Academic Search

Microsoft Academic Search, im Folgenden als *MAS* bezeichnet, ist ein Dienst von Microsoft Research. Es handelt sich um eine Suchmaschine für wissenschaftliche Publikationen und Autoren.

MAS erhält seine Daten größtenteils direkt von Herausgebern bezahlpflichtiger Inhalte. Diese werden in größeren Abständen in die Datenbank eingepflegt, wie den Updates⁴¹ zur Suchmaschine zu entnehmen ist. Wie Orduña-Malea et al. (2014) in ihrem Paper *Empirical Evidences in Citation-Based Search Engines: Is Microsoft Academic Search dead?* unter Zuhilfenahme anderer Arbeiten argumentieren, umfasst die Datenbank etwa 50 Millionen Publikationen. In dem Paper wird von den Autoren ebenfalls dargestellt, dass die Datenbanken von MAS seit 2011 kontinuierlich weniger neue Publikationen erhalten. Zum Ende des Jahres 2013 sei der Update-Prozess vollkommen zum Erliegen gekommen. Belege hierfür liefert MAS sogar selbst, durch die Visualisierung der Anzahl indizierter Publikationen. Ein deutlicher Rückgang derer ist in der Abbildung 3.21, sowie der Tabelle 3.2.3

⁴¹Updates und Inhalte von MAS <http://academic.research.microsoft.com/About/help.htm#5>

übereinstimmend mit den Erkenntnissen von Orduña-Malea et al. (2014) ab dem Jahr 2011 zu sehen. Es handelt sich dabei um einen Screenshot der Publikations-Trends in MAS⁴² vom August 2014, bzw. den quantitativen Daten des zugehörigen öffentlichen Datensatzes. Ende des Jahres 2013 wurde die eigenständige API von MAS abgeschaltet und stattdessen eine OData-API⁴³ über *Microsoft Azure Marketplace*⁴⁴ zur Verfügung gestellt.

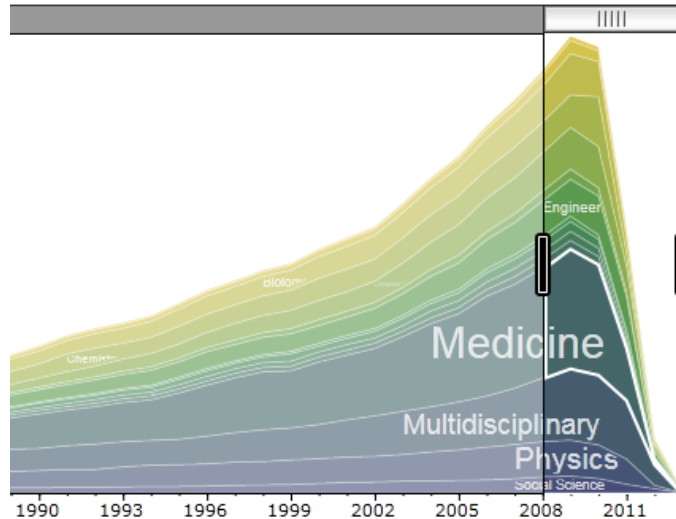


Abbildung 3.21.: Absolute Anzahl der in MAS gelisteten Publikationen pro Jahr, gruppiert nach Forschungsdisziplinen
Quelle: MAS Domain Trends⁴², selbst erstellter Screenshot

Jahr	Paper
2009	1.956.306
2010	1.869.993
2011	956.617
2012	239.459
2013	36.807
2014*	103.373

Tabelle 3.1.: Anzahl der in MAS gelisteten Publikationen pro Jahr von 2009 bis 2014.

* Bis einschließlich 08.09.2014

Datenquelle: Microsoft Academic Datensatz⁴⁴

Über die Suchmaske der Benutzeroberfläche können verschiedene Suchvorgänge gestartet werden. Elemente einer Boolean-Suche können verwendet werden. Es wird sich dabei an den gängigen Funktionen von großen Suchmaschinen wie Bing oder Google orientiert⁴¹. Die Übersichtsseite listet die Publikationen mit Titel und Teilen des Abstracts auf, falls

⁴²MAS Domain Trends <http://academic.research.microsoft.com/DomainTrend>

⁴³Open Data Protocol, entwickelt von Microsoft <http://www.odata.org/>

⁴⁴Microsoft Academic Datensatz <https://datamarket.azure.com/dataset/mrc/microsoftacademic>

vorhanden. Auch die Autoren und die Anzahl der Zitationen werden angegeben. Eine Sortierung der Einträge kann nur nach Microsofts nicht im Detail veröffentlichtem Ranking-Algorithmus vorgenommen werden.

Für Autoren der Datenbank existieren jeweils Profile mit weiteren Informationen über den Autor, allen in MAS enthaltenen Publikationen sowie Co-Autoren. Zusätzlich werden Aktivitäten des Autors bei Konferenzen und in Journalen gezeigt.

In MAS integriert sind verschiedene interaktive Möglichkeiten der Visualisierung. Es lässt sich für jeden Autor ein Co-Autorschafts-Graph und ein Zitations-Graph darstellen, sofern die notwendigen Daten verfügbar sind. Die Darstellung der Graphen erfordert die Installation des Microsoft-Silverlight-Plugins. Ein Beispiel eines solchen Graphen ist in der Abbildung 3.22 dargestellt. Interaktiv können weitere Autoren nachgeladen werden. Auch die Anzahl der angezeigten Autoren lässt sich variieren. In jedem Profil eines Autors ist eine histogrammatische Ansicht seiner Publikationen verfügbar. Diese können pro Jahr oder kumulativ angezeigt werden. Eine Schlagwortwolke über die wichtigsten thematischen Begriffe im Bezug auf den Autor ist ebenfalls auf jeder Profilseite zu sehen.

Werden statt der Co-Autorschafts-Beziehungen Zitations-Beziehungen zwischen Publikationen gezeigt, lassen sich interaktiv die Referenzen jeder Publikation nachladen. Das ermöglicht die explorative Darstellung einer beliebig langen Kette von Zitationen über diverse Publikationen hinweg.

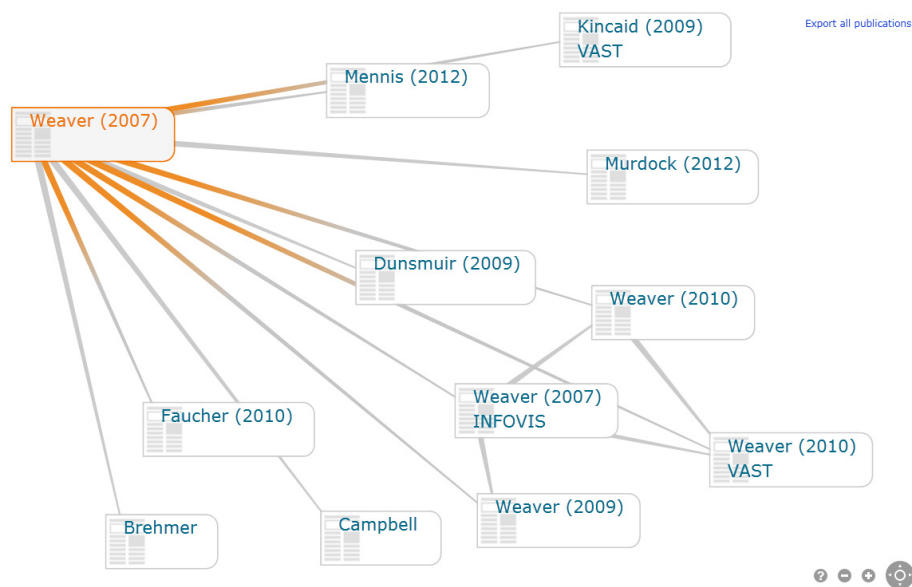
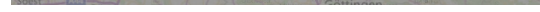


Abbildung 3.22.: Ein Zitations-Graph in MAS, die Publikation *Weaver (2007)* (links) wird von den restlichen 11 Publikationen zitiert

Quelle: MAS, Screenshot des Zitationsgraphen der Publikation

Eine weitere Visualisierung zeigt die aktuellen Trends in wissenschaftlichen Disziplinen⁴⁵. Die Darstellung nennt sich *Publication Trends*. Darin werden alle Publikationen

⁴⁵MAS Publication Trends <http://academic.research.microsoft.com/DomainTrend>



—

1000

Aktivitäten nach Ländern beziehungsweise Institutionen darzustellen. Ein Beispiel ist in der Abbildung 3.24 dargestellt. Auf diese Weise lassen sich besonders aktive Forschungsstandorte identifizieren. Die Darstellung ist interaktiv bedienbar. Zu jedem dargestellten Punkt können weitere Informationen angezeigt werden.

Insgesamt nutzt MAS das Potenzial an Visualisierungen seines Datenbestands gut. Es werden diverse Visualisierungen angeboten, die jedoch auf der Webseite leicht zu übersehen sind. Einige der Visualisierungen sind lediglich Prototypen, was den Nutzen aber nicht minimiert. Die Daten werden aktuell nachgeladen. Dies bietet einen großen Vorteil gegenüber „State-of-the-art“ Betrachtungen. Vorteilhaft ist auch die Praxis von MAS, die Datenbestände kostenfrei über eine API zur Verfügung zu stellen. Nach den geschilderten Beobachtungen der fehlenden Aktualisierungen des Datenbestandes von Orduña-Malea et al. (2014) schwindet jedoch leider der Nutzen dieser Datenbasis. Die Vermutung liegt nahe, dass die Finanzierung von MAS zurückgefahren wird und die Suchmaschine lediglich verwaltet wird. Auch ist die neu gestaltete API in ihren Funktionen unzureichend. Es ist derzeit (Juli 2014) noch keine Volltextsuche der Datensätze verfügbar, obwohl die neue API bereits seit Ende 2013 online ist. Dieser Umstand macht die Datensätze quasi nutzlos, sofern die knapp 100 Gigabyte an Daten⁴⁷ nicht manuell heruntergeladen und in einem eigenen Datenbankdienst verwendet werden. Die MAS API wird gesondert im Kapitel 4 bezüglich der Nutzbarkeit für AcaVis betrachtet.

3.2.4. CiteSeer

CiteSeer ist eine Online-Suchmaschine für wissenschaftliche Publikationen hauptsächlich aus den Bereichen der Informatik und Informationswissenschaften⁴⁸. Das Projekt wurde bereits im Jahr 1997 am NEC Research Institute in Princeton ins Leben gerufen und sollte zunächst nur einen prototypischen Charakter haben⁴⁹. Das gewachsene Nutzeraufkommen machte es den Entwicklern zufolge einige Jahre später notwendig, die Architektur komplett umzubauen, was in der momentan laufenden Version CiteSeerX Beta mündet.

Es handelt sich bei CiteSeer nicht um eine digitale Bibliothek von Publikationen sondern um eine echte Suchmaschine mit eigener Datensammlung durch einen Webcrawler⁵⁰. Der eingesetzte Webcrawler wird als selektiv beschrieben, beschränkt auf Webseiten, die in einer vorgegebenen Liste enthalten sind oder durch Benutzer zur Indizierung eingetragen wurden. Auf der Webseite⁵⁰ ist auch angemerkt, dass der Index bestehender Publikationen regelmäßig in nicht genannten Abständen aktualisiert wird.

⁴⁷MAS-Datensätze <https://datamarket.azure.com/dataset/mrc/microsoftacademic#schema>

⁴⁸Funktionen von CiteSeer <http://csxstatic.ist.psu.edu/about>

⁴⁹Historie von CiteSeer <http://csxstatic.ist.psu.edu/about/history>

⁵⁰Eigenschaften des Webcrawlers <http://csxstatic.ist.psu.edu/about/crawler>

Years of Citing Articles

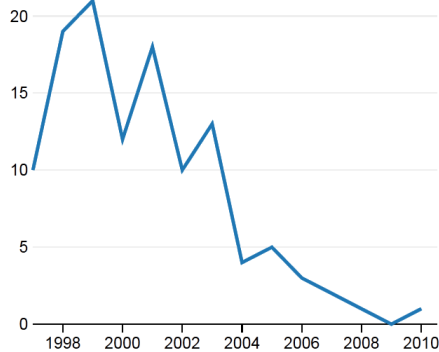


Abbildung 3.26.: Histogramm der Zitationen pro Jahr
Quelle: CiteSeer, Screenshot

Auf der Übersichtsseite⁴⁸ sind die einzelnen Fähigkeiten von CiteSeer aufgeführt. Die Kernfunktionen daraus sind im Folgenden aufgeführt. Der Webcrawler ist in der Lage, wissenschaftliche Publikationen auf den untersuchten Webseiten auffindig zu machen und sämtliche verfügbaren Informationen zu indizieren. Es werden PDF- und PostScript-Dateien analysiert. So werden automatisch Autoren, Titel und Abstract einer Publikation ausgelesen. Eine weitere wichtige Komponente ist die automatisierte Analyse des Literaturverzeichnisses jeder Arbeit. Die Verweise werden strukturiert in der Datenbank abgelegt und erlauben damit eine Zuordnung jeder Arbeit in den Zusammenhang mit anderen Publikationen und deren Autoren. Außerdem werden, sofern verfügbar, die Volltexte jeder Publikation analysiert, was eine inhaltliche Einordnung ermöglicht.

Die Suchmaske von CiteSeer erlaubt es, den gesamten Datenbestand zu durchsuchen. Die Sortierung der Ergebnisse kann anders als bei anderen Suchmaschinen selbst gewählt werden. Für jede Publikation in der Datenbank lassen sich alle gesammelten Informationen darstellen. Das Abstract und die zugehörige indizierte Webseite können eingesehen werden. Der Link zum Originaldokument steht ebenso zur Verfügung, wie die vollständige Liste der Verweise auf andere Publikationen. Sofern diese zitierten Publikationen ebenfalls in der CiteSeer-Datenbank enthalten sind, kann durch einen URL-Verweis darauf zugegriffen werden. Die einzige Form von Visualisierung der Daten aus CiteSeer ist durch ein Histogramm, wie in der Abbildung 3.26 zu sehen, gegeben. Dargestellt ist die Anzahl der Referenzen anderer Publikationen auf die vorliegende Publikation pro Jahr. Die Daten sind jeweils über das Jahr summiert und berücksichtigen ausschließlich die in CiteSeer indizierten Publikationen. Die Darstellung gibt einen Hinweis auf die Reichweite der Publikation und deren Bezug zur aktuellen Forschung.

Viele der gesammelten Daten von CiteSeer sind auch über eine öffentlich zugängliche API⁵¹ verfügbar. Der Webcrawler steht hier ebenfalls zur Verfügung und ermöglicht ein selbstständiges Durchsuchen von Webinhalten.

CiteSeer bietet einen überzeugenden unabhängigen Ansatz für die Recherche wissenschaftlicher Publikationen. Das Konzept ist in vielen Punkten angelehnt an die bewährten Verfahren der großen kommerziellen Konkurrenten Google Scholar und Microsoft Academic Search. Die Daten sind in CiteSeer aber insgesamt durchsichtiger, da sämtliche Informationen, inklusive der Literaturverzeichnisse, zur Verfügung stehen. Leider ist die Reichweite des Projekts durch die bereits genannten Aspekte und die begrenzten Ressourcen der Entwickler eingeschränkt. Auch das Potenzial an Visualisierungen zur Unterstützung der Recherchen wird nur minimal genutzt. Durch die Verfügbarkeit der Zitationen anderer

⁵¹Informationen zur CiteSeer-API <http://csxstatic.ist.psu.edu/about/data>

Arbeiten besteht aber prinzipiell die Möglichkeit, die Plattform durch aussagekräftige Visualisierungen anzureichern. Eine solche Untersuchung von Zitations-Netzwerken auf Basis der Daten von CiteSeer wurde zum Beispiel von Khalid et al. (2011) im Paper *Citation network visualization of CiteSeer dataset* praktisch durchgeführt.

3.2.5. Harzing's Publish or Perish

Publish or Perish ist ein desktopbasiertes Programm zur automatisierten Abfrage bibliografischer Daten aus Suchmaschinen für wissenschaftliche Literatur. Es wurde von Anne-Wil Harzing entwickelt. Das Programm ist kostenlos auf der Webseite⁵² von Anne-Wil Harzing verfügbar und wird regelmäßig weiterentwickelt. Die derzeitige Version unterstützt Anfragen an die Suchmaschinen Google Scholar und Microsoft Academic Search.

Die Suche nach Publikationen kann Autor-basiert, Journal-basiert und basierend auf der jeweiligen Standardsuche durchgeführt werden. Zurückgegeben wird eine Liste von Ergebnissen, die die ausgewählte Suchmaschine geliefert hat. Zusätzlich berechnet das Programm über zehn bibliografische Metriken⁵³ basierend auf der Liste von Publikationen. Die vollständige Liste lässt sich in verschiedenen Dateiformaten, darunter BibTex, ISI und CSV, exportieren. So lassen sich die gesammelten Daten mit vielen der im Abschnitt 3.1 vorgestellten Programme zur Visualisierung nutzen.

Das Programm ignoriert die Aufforderung von Google Scholar⁵⁴, das automatisierte Abfragen von Ergebnissen zu unterlassen. Die Anzahl möglicher Abfragen für einen gewissen Zeitraum und die zurückgelieferten Ergebnisse werden von Google Scholar stark begrenzt. Im normalen Gebrauch ist dies nicht relevant. Lediglich Webcrawler wie Publish or Perish werden diese Grenze erreichen. Wie die Autorin von Publish or Perish in der Dokumentation⁵⁵ angibt, wurde deshalb eine „Anfragebremse“ in das Programm integriert. Diese soll das Sperren der IP-Adresse durch Google Scholar verhindern.

Publish or Perish ist ein interessanter Ansatz der automatisierten Abfrage von bibliografischen Daten aus den gängigen Suchmaschinen für wissenschaftliche Publikationen. Auch wenn die Gefahr einer temporären Sperrung der eigenen IP-Adresse durch die Suchmaschinen droht, ist das Programm für den privaten Gebrauch gut geeignet. Die zur Verfügung stehenden Exportformate lassen eine Benutzung der Daten mit einigen anderen vorgestellten Programmen der Analyse und Visualisierung wie zum Beispiel VOSviewer, CitNetExplorer und Sci²Tool zu.

⁵²Publish or Perish <http://www.harzing.com/pop.htm>

⁵³Metriken in Publish or Perish <http://www.harzing.com/pophelp/metrics.htm>

⁵⁴Hinweis auf das Verbot von Webcrawlern <http://scholar.google.com/intl/en/scholar/help.html#export>

⁵⁵„Anfragebremse“ in Publish or Perish <http://www.harzing.com/pophelp/faq.htm#Q1010>

3.3. Das fehlende Puzzlestück

Die in diesem Kapitel dargestellten Systeme zur bibliografischen Organisation und Visualisierung sind sehr vielfältig. Zu jedem der drei großen Teilbereiche Organisation, Visualisierung und Datensammlung sind bereits verschiedene Ansätze mit den jeweils diskutierten Vor- und Nachteilen vorhanden. Die Darstellung erhebt keinerlei Anspruch darauf, wirklich jeden existierenden praktischen Ansatz berücksichtigt zu haben. So lassen sich im Rahmen dieser Zusammenfassung von praktischen Ansätzen einige Erkenntnisse über die besonders gelungenen Ansätze und über vorhandene Defizite sammeln.

Im Abschnitt 3.1 wurden zu Beginn dieses Kapitels diverse Programme zur vorteilhaften Aufarbeitung bibliografischer Daten vorgestellt. Ähnliche methodische Ansätze, insbesondere der Netzwerkanalyse, stellen sich hier als besonders mächtig heraus, gerade im Umgang mit großen Datenmengen. Dennoch bestehen einige deutliche Defizite:

1. Alle vorgestellten Programme, abgesehen vom Alluvial Generator, sind desktopbasiert. Wünschenswert wäre aber eine webbasierte Anwendung, die die Nutzbarkeit über verschiedene Betriebssysteme und Endgeräte hinweg zu ermöglicht.
2. Die Arbeitsabläufe zur Erstellung einer aussagekräftigen Visualisierung sind vielfach kompliziert. Anzustreben sind bestimmte Typen von Standardgrafiken, die eine optisch und inhaltlich ansprechend vorbereitete Visualisierung für genau einen Zweck liefern. Dies würde die Visualisierungen einem breiten Publikum zugänglich machen und weniger Fachkenntnis erfordern.
3. Keines der vorgestellten Programme unterstützt eine Beschaffung wirklich aussagekräftiger Daten. Die wenigen vorhandenen integrierten Datenquellen liefern lediglich fachspezifische Informationen oder unzureichende Datensätze, beispielsweise ohne Zitationen und Referenzen. Vor dem Hintergrund, dass viele der Programme eine Visualisierung von Netzwerken anstreben, ist das Fehlen von entsprechenden Datenquellen unvorteilhaft. Angebracht wäre eine automatisierte Abfrage dieser Netzwerkinformationen, um das Potenzial der zugehörigen Visualisierungen nutzen zu können.

Ergänzend wurden im Abschnitt 3.2 die bibliografischen Organisations- und Informationssysteme betrachtet. Auch diese bieten Potenzial zur Verbesserung:

1. Google Scholar und CiteSeer nutzen die Mittel der Visualisierung zur Unterstützung der Recherche von wissenschaftlichen Publikationen kaum. Gerade Google Scholar stehen die finanziellen Möglichkeiten und die notwendige Datenbasis zur Verfügung, um Visualisierungen zu integrieren.
2. Google Scholar und MAS bieten außerdem keine Möglichkeiten, eine individuelle bibliografische Literaturverwaltung online vorzunehmen.
3. Die Bereitstellung von Daten über APIs ist äußerst beschränkt. Die Daten werden entweder garnicht oder nur in Teilen publiziert. Zu den zurückgehaltenen Daten gehören leider gerade jene, wie Zitationen und Referenzen, die von besonderem Interesse sind. Im Fall von MAS werden diese lobenswerter Weise bereitgestellt, die API ist aber ohne Suchfunktion und ohne zukünftige Updates der Datenbank kaum nutzbar.

Die aufgeführten Punkte sind jedoch nicht ausschließlich als Kritik an den vorgestellten Programmen zu verstehen, sondern sollen auch wichtige Anregungen für die Entwicklung von AcaVis sein. Ziel von AcaVis ist es, die Funktionen in einer Applikation zu bündeln und systemübergreifend mittels einer Weboberfläche anzubieten. Die Formulierung der Anwendungsfälle für AcaVis muss wegen der in diesem Kapitel gesammelten Erkenntnisse zweigeteilt werden. Auf der einen Seite stehen Anwendungsfälle, deren Umsetzbarkeit lediglich durch den zu leistenden Arbeitsaufwand begründet ist. Auf der anderen Seite stehen Anwendungsfälle, deren Notwendigkeit zum Beispiel aus diesem Kapitel folgt, die aber aus verschiedenen Gründen praktisch nicht mit angemessenem Aufwand umsetzbar sind.

Die praktische Machbarkeit ist in diesem Zusammenhang insbesondere durch das Vorhandensein einer zugänglichen Datenbasis bedingt. Die ursächlichen Defizite wurden bereits besprochen. Dieser Eckpfeiler der späteren Implementierung wird deshalb noch vor den Anwendungsfällen im nachfolgenden Kapitel 4 erläutert. Daraus folgt schließlich die Zerteilung der Anwendungsfälle in die genannten Kategorien.

4. Schnittstellen zu externen Informationssystemen

Grundlage einer jeden Visualisierung sind die Daten, die visualisiert werden sollen. Visualisierungen dienen, wie im Kapitel 2 ausführlich erläutert, nicht dem Selbstzweck. Im Bezug auf das Ziel dieser Arbeit sind grundsätzlich sämtliche Daten zu wissenschaftlichen Publikationen relevant. Das umfasst Metadaten und Daten zur Identifikation, wie auch die Publikationen im Volltext und natürlich Informationen, die sich aus den Publikationen generieren lassen. Gegenüber der Relevanz steht die Verfügbarkeit der Daten. Trotz der heutzutage weitreichenden Möglichkeiten stehen die Informationen vielfach nicht einheitlich oder nur gegen ein Entgelt zur Verfügung. Die wissenschaftlichen Prinzipien der gegenseitigen Kontrolle und des kooperierenden Arbeitens erfordern im Grunde aber eine umfängliche und öffentliche Verfügbarkeit aller Informationen.

Vielen Hochschulangehörigen stehen die nicht-öffentlichen Publikationen in den meisten Fällen trotzdem über hochschuleigene Zugänge zu kostenpflichtigen Verzeichnissen zur Verfügung. Nachteilig ist hierbei die Vielzahl verschiedener Verzeichnisse von Publikationen. Große Verlage wissenschaftlicher Publikationen pflegen in der Regel nur Verzeichnisse ihrer eigenen Veröffentlichungen und erlauben den automatisierten Zugriff wiederum nur gegen weiteres Entgelt. Diese Praxis führt einerseits dazu, dass nur wenige eigenständige und übergreifende Verzeichnisse existieren. Andererseits muss der Zugriff finanziert werden, sodass monetäre Interessen in die Verbreitung der gebündelten Informationen hineinspielen. Einige solcher übergreifenden Verzeichnisse wurden im vorangegangenen Kapitel 3 mit Fokus auf die vorhandenen Visualisierungen vorgestellt. Wie in der dortigen Zusammenfassung beschrieben, ist jedoch keines dieser Systeme prädestiniert, die genannten Eckpunkte ohne Einschränkung zu erfüllen.

Der Anspruch von AcaVis ist es, die vielen positiven Aspekte existierender praktischer Ansätze im Rahmen der Aufgabenstellung zu bündeln und gegebenenfalls zu ergänzen. Demnach läuft die Entwicklung von AcaVis auf die Entwicklung eines eigenständigen webbasierten Systems hinaus. Es ist im Rahmen dieser Arbeit außerdem nicht praktikabel, eine eigene Publikationsdatenbank aufzubauen. Einerseits begründet durch die zuvor genannten Kosten und andererseits durch die enorme Größe der Datensätze von mehreren hundert Gigabyte, nach Darstellungen von Orduña-Malea et al. (2014). Daher bietet sich die Lösung an, zugängliche APIs zu existierenden gebündelten Verzeichnissen zu verwenden. Im nächsten Abschnitt werden deshalb die bereits vorgestellten Verzeichnisse kurz hinsichtlich ihrer Verwendbarkeit als Datenbasis für AcaVis betrachtet.

4.1. Zusammenfassung der geeigneten Informationssysteme

Betrachtet werden die im Abschnitt 3.2 vorgestellten Verzeichnisse: Google Scholar, CiteSeer, Mendeley und MAS. Das Programm Harzing's Publish or Perish ist in diesem Sinn kein Verzeichnis für Publikationen, sondern lediglich ein desktopbasierter Webcrawler und wird aus diesem Grund nicht betrachtet.

Google Scholar

Google Scholar ist als Datenbasis für AcaVis nicht geeignet, weil keine öffentlich zugängliche API zur Verfügung gestellt wird und die Anwendung von Webcrawlern unerwünscht ist. Weiter wird die Zahl an Zugriffen pro Zeiteinheit limitiert, was den Einsatz eines Webcrawlers unter Missachtung des entsprechenden Verbots auch nahezu unmöglich macht. Das Problem und entsprechende Lösungsansätze wurden bereits im Zusammenhang mit dem Programm Harzing's Publish or Perish im Abschnitt 3.2.5 erläutert.

CiteSeer

Die Publikations-Datenbank von CiteSeer wird nach Angaben der Entwickler nur durch einen Webcrawler und durch Aktivitäten der Community aufgebaut. Somit ist sie unabhängig von kostenpflichtigen Datensätzen der Verlage. Der Zugang zur Datenbank ist unentgeltlich über eine öffentlich zugängliche API möglich, wobei jedoch nicht alle Daten zur Verfügung stehen. Eine negative Folge der Beschaffenheit der Datenbank ist es aber, dass diese viele wichtige Publikationen vermissen lässt. Die kollaborativen Einträge der Community können die Menge an Daten, die durch kostenpflichtige Datenzugänge verfügbar wäre, nicht vollständig kompensieren. Der Webcrawler selbst ist nur in der Lage, Publikationen auf öffentlichen Internetseiten von Autoren und Instituten zu indizieren, nicht aber auf solchen Internetseiten für die eine Autorisierung erforderlich ist. Ein weiteres Defizit konzeptioneller Art von CiteSeer ist es, dass der Fokus auf Publikationen im technischen Bereich und hier fast ausschließlich auf Publikationen der Informatik und Informationswissenschaften liegt. Die Verwendbarkeit der CiteSeer API in AcaVis ist also möglich, wird aber nicht verfolgt.

Mendeley

Mendeley ist in den Ursprüngen ein System zur Verwaltung von Sammlungen von Publikationen, sowie zur Annotation und Organisation dieser. Hinzugekommen ist der Mendeley-Katalog, welcher Informationen zu einer Vielzahl an Publikationen enthält. Die Informationen werden durch Aggregation der Publikations-Sammlungen von Mendeley-Benutzern erstellt⁵⁶. Das Verzeichnis ist somit unabhängig von kostenpflichtigen Zugängen zu verlagseigenen Verzeichnissen. Es kann außerdem angenommen werden, dass das Verzeichnis vergleichsweise umfangreich ist, da die Informationen von den Mendeley-Benutzern primär

⁵⁶Datensammlung des Mendeley-Katalogs <http://blog.mendeley.com/tag/mendeley-web-catalog/>

zum Selbstzweck erstellt werden. Die übergreifende Wertschöpfung aus diesen Daten geschieht sekundär seitens Mendeley in Form des Mendeley-Katalogs. Es steht zudem eine API zur Datenbank des Mendeley-Katalogs zur Verfügung. Der Zugang zur API erfordert eine kostenlose Registrierung⁵⁷. Die API ist sehr gut dokumentiert und stellt sämtliche im Mendeley-Katalog enthaltenen Informationen zur Verfügung⁵⁸.

Die Datenbank von Mendeley enthält einige Besonderheiten. Es sind keine Listen von Referenzen und Zitationen der Publikationen abrufbar, weder im Mendeley-Verzeichnis, noch über die API. Dies ist ein Defizit, da Referenzen und Zitationen sehr häufig zu Analysen verschiedenster Art von Publikationen verwendet werden. Eine Besonderheit die sonst in keinem der betrachteten Verzeichnisse anzutreffen ist, ist die Verfügbarkeit von Informationen über die Benutzer, die diese Publikationen in ihrer Sammlung haben. Die Gesamtzahl dieser Benutzer ist jeweils aufgeschlüsselt in drei Kategorien. Benutzer nach wissenschaftlicher Disziplin, Benutzer nach akademischem Grad und Benutzer nach Staatsangehörigkeit. Alle weiteren benutzerspezifischen Daten sind anonymisiert. Die Verfügbarkeit dieser Informationen eröffnet interessante Möglichkeiten der weiterführenden Analyse. Trotz der Relevanz für das Thema dieser Arbeit werden diese in AcaVis vorerst nicht weiterverfolgt. Ansätze einer dahingehenden Arbeit könnten zum Beispiel die Kategorisierung von Publikationen, die Ermittlung der Relevanz oder die geografische Verbreitung sein. Die folgende, nicht abschließende Liste enthält allgemeine Aspekte, die in diesem Zusammenhang mindestens geklärt werden sollten:

- Ist die Integrität der Daten gewährleistet und wie beeinflussen falsche Informationen die Aussagekraft?
- Ist die Betrachtung des Benutzerspektrums von Mendeley disziplinär/interdisziplinär repräsentativ?
- Hat die Darstellung der Informationen einen Wert, wenn ja welchen?
- Können weitere „versteckte“ Informationen mittels Daten-Mining gewonnen werden, wenn ja welche?

Microsoft Academic Search

Microsoft Research stellt im Gegensatz zum Konkurrenten Google eine API zu seiner Suchmaschine MAS für wissenschaftliche Publikationen zur Verfügung⁵⁹. Es lassen sich über diese API nahezu alle über das MAS-Verzeichnis verfügbaren Informationen abfragen. Es bestehen keine Limitierungen bezüglich der Häufigkeit von Anfragen oder der Anzahl zurückgelieferter Ergebnisse. Insbesondere sind Informationen über Zitationen und Referenzen zwischen den Publikationen verfügbar, was die Datengrundlage für die sehr verbreiteten Analysen von Autoren- und Publikations-Netzwerken liefert. Diese Verfügbarkeit verleiht der MAS-Datenbank einen besonderen Stellenwert.

⁵⁷Mendeley-Catalog API <http://dev.mendeley.com/>

⁵⁸Mendeley-Catalog API Dokumentation https://api.mendeley.com/apidocs#!/catalog_0/search

⁵⁹MAS Datensätze in Microsoft Azure Marketplace <https://datamarket.azure.com/dataset/mrc/microsoftacademic>

Trotz der deutlichen Vorteile muss eine Verwendung der MAS-Datenbank kritisch betrachtet werden. Die im Abschnitt 3.2.3 zusammengefassten Ergebnisse von Orduña-Malea et al. (2014) legen nahe, dass die zukünftige Pflege der MAS-Datenbanken mindestens infrage steht. So muss bei einer Verwendung von MAS als Datenbasis deutlich auf diesen Umstand hingewiesen werden.

Resümee

Google Scholar ist derzeit als Bestandteil von AcaVis aufgrund der fehlenden API ungeeignet. Sollte sich dies ändern so ist Google Scholar für diesen Zweck definitiv in Erwägung zu ziehen, da das Verzeichnis besonders umfangreich ist (Orduña-Malea et al., 2014) und gegenseitige Zitationen der Publikationen enthält. CiteSeer ist durch die Verfügbarkeit einer öffentlich zugänglichen API zwar grundsätzlich für eine Verwendung mit AcaVis geeignet, Qualität und Quantität der Daten bewegen sich aber auf einem niedrigeren Niveau als die der konkurrierenden Verzeichnisse. Darüber hinaus werden keine Informationen zu Referenzen und Zitationen bereitgestellt, welche für die besonders aussagekräftigen Netzwerk-Visualisierungen notwendig sind. Mendeley ist durch die verfügbare API zum Mendeley-Katalog ebenfalls geeignet, besonders da sehr reichhaltige Informationen zur Verfügung stehen. Leider werden durch Mendeley ebenfalls keine Referenzen und Zitationen bereitgestellt, mit den bereits genannten Folgen für AcaVis. Eine mögliche Substitution durch Benutzeranalysen wurde zuvor in den Grundzügen beschrieben. Die zuletzt betrachtete Datenbank von MAS ist besonders reichhaltig hinsichtlich der dargestellten Informationen. Die API ist öffentlich, die Benutzung keinerlei schwerwiegenden Einschränkungen unterlegen. Trotz der Frage nach dem Fortbestand und der Pflege des MAS Datenbestands ist diese Datenquelle aufgrund der genannten Punkte zu bevorzugen.

Zukünftige äußere Einflüsse, wie eine Abschaltung von MAS oder eine Aufwertung der anderen Systeme durch zusätzliche Datensätze oder eine API, sollten handhabbar sein, da sie nicht ausgeschlossen werden können.

4.2. Die API von Microsoft Academic Search

Die derzeit aktuelle API von MAS löste Ende 2013 die bis dahin bestehende API ab⁶⁰. Während die alte API zur Benutzung einer individuellen Freischaltung durch Microsoft Research bedurfte, ist der Zugang zur neuen API ohne Beschränkungen möglich. Voraussetzung ist eine kostenlose Anmeldung im Microsoft Azure Marketplace⁵⁹. Die technische Grundlage der aktuellen MAS API bildet eine Schnittstelle im OData-Format⁶¹, einem von Microsoft entwickelten Protokoll für Datenbankabfragen über das HTTP-Protokoll. Die MAS API verwendet das OData-Protokoll der Version 2, es steht jedoch nur ein eingeschränkter⁶² Befehlssatz zur Verfügung. Ergebnisse werden in den Formaten JSON und XML ausgeliefert, wobei erstere Variante bei gleicher Semantik weniger syntaktischen Overhead liefert.

⁶⁰Ankündigung des Wechsels der MAS API <http://social.microsoft.com/Forums/de-DE/f4ec949e-d0cf-4f37-a1c8-b50e2ba00c30/>

⁶¹OData Webseite <http://www.odata.org/>

⁶²OData-Unterstützung <http://msdn.microsoft.com/en-us/library/gg312156.aspx>

Einschränkungen

Die bereitgestellte API besitzt einige schwerwiegende Einschränkungen. Es ist seit der Umstellung der API nicht mehr möglich, eine Volltextsuche nach Publikationen durchzuführen. Auch das Durchsuchen einzelner Attribute wie dem Titel oder dem Abstract ist nicht mehr möglich, ebenso eine Suche nach anderen identifizierenden Attributen wie der ISBN-Nummer oder dem DOI-Identifikator. Diese Einschränkung macht die MAS Datenbank alleine nahezu unbrauchbar.

Weitere Einschränkungen betreffen die technische Umsetzung der API. Wie erwähnt stehen nicht alle Funktionen zur Verfügung, die in der OData-Spezifikation⁶³ enthalten sind. Es fehlt insbesondere eine Umsetzung der Funktion *expand*, welche gleichzusetzen ist mit der Join-Funktion regulärer Datenbanken. Im Ergebnis müssten notwendige Join-Befehle anderweitig nachgebildet werden. Die zweite Einschränkung betrifft die Anfragen an die API. Das OData-Protokoll baut auf dem verbreiteten HTTP-Protokoll auf. OData-Anfragen werden über den HTTP-Header verschickt, welcher aus Sicherheitsgründen einer Größenbeschränkung unterliegt. Überschreitet eine OData-Anfrage die im Webserver der API definierte Beschränkung, wird die Anfrage nicht ausgeführt und stattdessen ein HTTP-Fehler *413 Entity Too Large* zurückgegeben. Die Beschränkung ist in jedem Webserver individuell festgelegt und bewegt sich üblicherweise im Bereich von vier bis 16 Kilobyte. Deutlich kleinere oder größere Werte sind aber prinzipiell möglich. Dieses Verhalten zeigt auch der Webserver der MAS API.

Außerdem sind die gelieferten Datensätze teilweise inkohärent bezüglich Fremdschlüsselbeziehungen. Es sind Fremdschlüssel auf Entitäten enthalten, die in der Datenbank nicht existieren. Mutmaßlich eine Ursache der nicht verfügbaren Join-Funktion. Eine direkte Verarbeitung der Daten ist daher nicht ohne Weiteres möglich.

Datenbankschema

Nach der Betrachtung der sekundären Eigenschaften der API wird nun das Schema der zugrundeliegenden Datenbank vorgestellt. Die Informationen über das Datenbankschema entstammen der tabellarischen Darstellung auf der Übersichtsseite des API-Dienstes in Microsoft Azure Marketplace⁶⁴. Die tabellarischen Ausführungen wurden in ein erweitertes Entity-Relationship-Modell umgewandelt, welches in der Abbildung 4.1 zu sehen ist. Die dargestellten Beziehungen und Constraints werden in der Quelle nicht explizit genannt und wurden nach logischen Gesichtspunkten und implizierenden Bezeichnungen ergänzt.

Die Darstellung berücksichtigt nur die zunächst für AcaVis relevanten Tabellen. Die restlichen Tabellen, sowie alle verfügbaren Tabellen-Attribute können den Beschreibungen auf Microsoft Azure Marketplace⁶⁴ entnommen werden. Die Umsetzung des vorliegenden Modells wird durch eine Transformation in ein relationales Datenbankschema mittels Entitäten und deren Beziehungen vorgenommen. Genauer: Für alle orange gefärbten

⁶³ODataV2 Spezifikation <http://www.odata.org/documentation/odata-version-2-0/uri-conventions>

⁶⁴Modell der Datenbank, die MAS zugrundeliegt <https://datamarket.azure.com/dataset/mrc/microsoftacademic#schema>

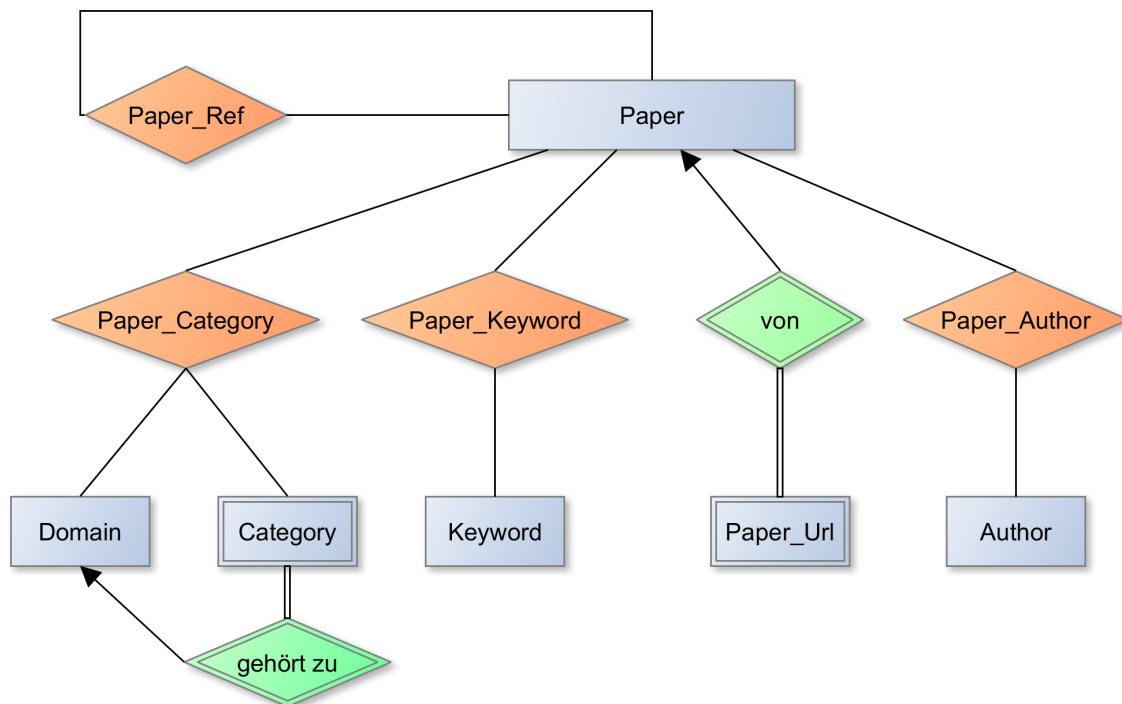


Abbildung 4.1.: Erweitertes Entity-Relationship-Modell der MAS zugrundeliegenden Datenbank

Datenquelle: Microsoft Azure Marketplace⁶⁴

Visualisierung und Aufarbeitung: eigenes Werk

Beziehungen existieren auch Tabellen in der Datenbank von MAS, welche die jeweilige Ausprägung der Beziehung durch deren identifizierende Schlüssel verwalten. Die grün gefärbten Beziehungen, namentlich *von* und *gehört zu*, existieren dagegen nicht als Tabellen in der Datenbank. Die Entität *Paper_Url* ist alleiniger Eigentümer der Beziehung, sodass die Ausprägung innerhalb dieser Entität vermerkt werden kann und nicht in eine zusätzliche Tabelle der Beziehung *von* ausgelagert werden muss. Die Beziehung *gehört zu* definiert lediglich einen logischen Constraint, nämlich dass eine Kategorie (Subdisziplin) ohne eine Domain (Disziplin) nicht existieren kann. Aus diesem Grund ist auch hier keine eigene Tabelle notwendig.

Maßnahmen zur Korrektur der Defizite von MAS

Um die Verwendbarkeit der MAS API für AcaVis zu sichern, werden die hier beschriebenen Maßnahmen ergriffen, die die vorhandenen Beschränkungen und Defizite handhabbar machen:

1. Die fehlende Suchfunktion über die MAS API wird durch einen Web-Crawler ergänzt, welcher sich der regulären Webseite von MAS zur Suche bedient. Somit stehen

alle dort verfügbaren Daten und Funktionen der Suche theoretisch auch in AcaVis zur Verfügung. Eine Weiterverarbeitung der Suchergebnisse kann mit Hilfe des MAS-eigenen Primärschlüssels für Publikationen und Autoren geschehen. Somit ist eine Verknüpfung der unvollständigen Informationen der Suchergebnisse mit den vollständigen Informationen der API möglich. Web-Crawling wird ausschließlich in Fällen benutzt, in denen die Funktionen dringend notwendig, aber mit Hilfe der API nicht realisierbar sind.

2. Das bereits beschriebene Fehlen der Join-Funktion von Tabellen in der MAS API ist wegen der Struktur des vorliegenden Datenmodells in zehn einzelnen Tabellen von klarer Bedeutung. Die Funktion wird entsprechend lokal implementiert, durch die Verwendung einer sogenannten In-Memory-Datenbank. Alle beteiligten Datensätze einer Abfrage in AcaVis werden zunächst aus der MAS API abgefragt, anschließend lokal in die temporäre In-Memory-Datenbank eingefügt und sind dann mit Hilfe einer vollständigen Abfragesprache, inklusive Join-Funktionalität, wie gewünscht abrufbar.
3. Die Inkonsistenz der Datenbestände von MAS wird ebenfalls durch die temporäre In-Memory-Datenbank korrigiert. Bestimmte Einstellungen erlauben es, diese inkonsistenten, im Fachjargon als *detached* bezeichneten, Datensätze bei der Überprüfung von zugehörigen Constraints zu ignorieren. Die Folge ist dann aber, dass diese Datensätze nicht zur Ausgabe zur Verfügung stehen.
4. Bezüglich der Größenbeschränkung von HTTP-Headern wird folgende Lösung angestrebt: Es wird eine maximale Länge (in Byte) einer einzelnen Anfrage definiert. Anfragen einer Größe über diesem Limit werden in einzelne Anfragen gültiger Länge aufgeteilt und sukzessiv durchgeführt. Unter Umständen ist auch eine parallele Verarbeitung möglich, was eine Beschleunigung der gesamten Anfrage zur Folge hätte. Da nicht ausgeschlossen werden kann, dass sich die obere Grenze der Länge in Zukunft ändert, wird diese Option als konfigurierbarer Parameter ausgeführt.

Wie im Abschnitt 4.1 dargelegt, ist im Grunde jedes der betrachteten Informationssysteme mit Defiziten behaftet. Es wäre also wünschenswert, wenn zukünftige Änderungen in MAS oder ein kompletter Wechsel des AcaVis zugrundeliegenden Informationssystems möglich wäre. Daher wird für AcaVis eine Programmarchitektur angestrebt, die eine gute Wartbarkeit und Wiederverwendbarkeit durch folgende Maßnahmen ermöglicht: Erstens sollen Informationssysteme entkoppelt von der AcaVis-Oberfläche sein, zweitens der Zugriff auf ein implementiertes Informationssystem mit wenig Aufwand angepasst werden können und drittens soll es möglich sein, mehrere verschiedene Informationssysteme selektiv zu verwenden. Der zugehörige Entwurf ist dem Kapitel 6 zu entnehmen.

Nachdem also die prinzipielle Durchführbarkeit des Vorhabens durch die Verfügbarkeit einer hinreichenden Datenbasis und deren Anpassungen dargelegt ist, können auf dieser Basis die Anwendungsfälle entsprechend dem eingangs beschriebenen, zweigeteilten Schema formuliert werden.

5. Anwendungsfälle im Rahmen der Exploration und Visualisierung

Die Exploration von Publikationen im Sinne der Recherche ist seit jeher Teil der wissenschaftlichen Arbeit. So haben sich diverse Verfahrensweisen und Empfehlungen etabliert, wie eine Recherche von wissenschaftlicher Literatur erfolgen kann. Interessierten werden dazu Seminare und Kurse an den örtlichen Bibliotheken angeboten und es ist eine Vielzahl an Literatur zu diesem Thema zu finden.

Vor dem Hintergrund, dass sich die Art und Weise der Recherche durch die zunehmende Verwendung digitaler Informationssysteme im letzten Jahrzehnt deutlich verändert hat, sind heute einige Grundprinzipien bei der Recherche als essenziell zu betrachten. Die besondere Herausforderung besteht nicht mehr in der Beschaffung von Material, sondern in der treffenden Auswahl des Materials und Sicherung der Integrität von Informationen.

Einen Einblick in den typischen Ablauf der Recherche wissenschaftlicher Publikationen mit Hilfe von digitalen Informationssystemen geben zum Beispiel Hull et al. (2008) in ihrem Journalartikel *Defrosting the Digital Library: Bibliographic Tools for the Next Generation Web*. Ein Strukturdiagramm des Ablaufs wurde ebenfalls in diesem Artikel veröffentlicht und ist in der Abbildung 5.1 zu sehen.

Der Einstiegspunkt einer jeden Recherche ist demnach die Suche (Search), das „Stöbern“ (Browse) oder die Empfehlung (Recommend) einer Menge von Publikationen. Hull et al. (2008) unterteilen außerdem die einzelnen Aufgaben in jene die fast ausschließlich von Menschen erledigt werden (weiß) und jene, die überwiegend maschinell erledigt werden (blau). Natürlich unterliegen diese Aufgaben einem Wandel und die Kategorisierung ist heute nicht mehr unbedingt vollständig zutreffend. Insbesondere die Empfehlung von Publikationen erfolgt heute sowohl maschinell, als auch persönlich. Nach Hull et al. (2008) werden die Publikationen im Weiteren ausgewählt (Select), beschafft (Get), gespeichert (Save) und schließlich gelesen (Read). Diese als relevant eingestuften Publikationen fließen dann ggf. in die eigene Arbeit ein oder werden auf andere Weise nutzbar gemacht.

Der Anspruch von AcaVis ist es, ein Informationssystem bereitzustellen, welches bei der Bewältigung zum Beispiel dieser üblichen Aufgaben behilflich ist und diese mit Visualisierungen bereichert. Daher sind alle blau gefärbten Aufgaben aus der Abbildung 5.1 als Anwendungsfälle für AcaVis relevant. Die Aufgaben werden in drei Teilbereichen zusammengefasst: Der Suche von Publikationen, der Betrachtung von Daten und Metadaten einer Publikation und dem Anlegen von individuellen Sammlungen an Publikationen. Diese Anwendungsfälle werden in diesem Kapitel in den späteren Abschnitten *Suchfunktion*, *Detailansicht* und *Sammlungen von Publikationen* eingehender definiert. Zunächst sollen aber die in AcaVis vorgesehenen Visualisierungen und ihr jeweiliger Nutzen für die Recherche vorgestellt werden.

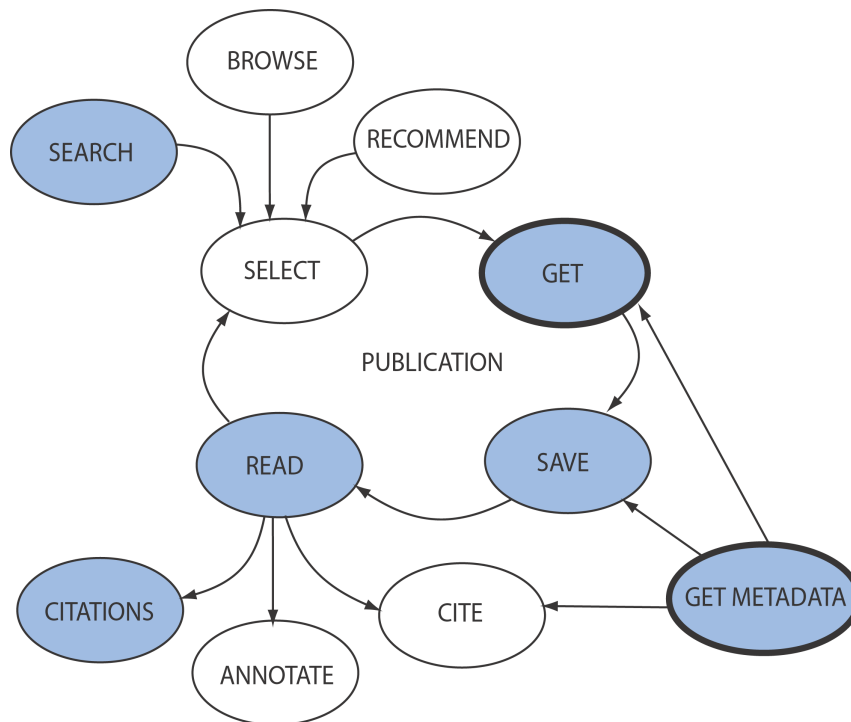


Abbildung 5.1.: Ein typischer Arbeitsablauf bei der Recherche von wissenschaftlichen Publikationen. Aufgaben in blau können und werden überwiegend maschinell erledigt, während Aufgaben in weiß fast nur von Menschen erledigt werden.

Quelle: Hull et al. (2008)

5.1. Visualisierungen

Die Visualisierungen, als zentrales Element von AcaVis, werden für sich betrachtet. Sie stellen erstens eine eigene Art des Anwendungsfalls dar und tauchen zweitens in ähnlicher Variante in verschiedenen Anwendungsfällen auf. Die hier aufgelisteten Visualisierungen sind diejenigen, die in AcaVis zur Umsetzung kommen werden. Die Umsetzung erfolgt mit Hilfe der JavaScript-Bibliothek *D3.js*⁶⁵. An entsprechender Stelle in den weiteren Abschnitten dieses Kapitels wird dann auf die dort verwendeten Visualisierungen verwiesen.

5.1.1. Verteilung von Publikationen über die Zeit

Die Visualisierung beinhaltet ein Balkendiagramm, welches eine Menge von Publikationen bezüglich ihrer Verteilung pro Jahr darstellt. Verwendet wird dafür das Erscheinungsjahr. Das zugrundeliegende Koordinatensystem beinhaltet als Abszissenachse einen Zeitbereich, unterteilt in Jahren, und als Ordinatenachse eine Anzahl von Publikationen. Ein Beispiel

⁶⁵Data-Driven Documents (D3.js) <http://d3js.org/>

einer solchen Visualisierung ist der Abbildung 5.2 zu entnehmen. Es handelt sich um einen Prototypen mit einem Datensatz von 98 Publikationen zum Thema „Citation Networks“.

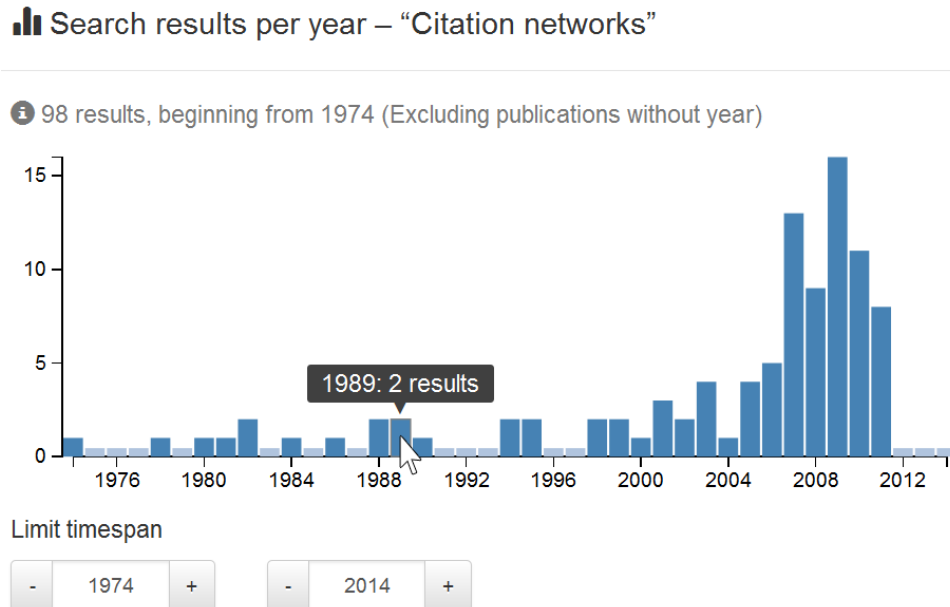


Abbildung 5.2.: Prototyp eines Balkendiagramms der Verteilung von Publikationen pro Jahr. 98 Publikationen aus einem Suchergebnis mit der Eingabe „Citation Networks“.

Quelle: eigenes Werk (Prototyp, Screenshot)

Folgende Erkenntnisse soll die Visualisierung einer Menge von Publikationen über die Zeit liefern:

1. Zu welchem Zeitpunkt wurden viele Publikationen aus der Menge publiziert?
2. Gibt es Zeitpunkte einer starken Änderung?
3. Gab es in der Vergangenheit Trends bzw. lässt sich ein gegenwärtiger Trend der Zahl an Publikationen erkennen?
4. Wurden die Publikationen kürzlich veröffentlicht, sind also Gegenstand aktueller Forschung, oder liegen die Veröffentlichungen länger zurück?

Folgende Möglichkeiten der Interaktion mit der Visualisierung bestehen:

1. Die Menge von Publikationen lässt sich bezüglich der darzustellenden Zeitspanne filtern.
2. Mausinteraktion mit den Datenbalken hebt deren exakte Werte in einem Tooltip hervor. Nützlich ist diese Funktion, da die Beschriftungen der Abszissenachse bei großen Datenmengen nicht mehr jährlich, sondern in größeren Abständen vorgenommen werden. Auch die Ordinatenachse ist nur in bestimmten Schritten beschriftet.

5.1.2. Zitationsnetzwerk mit Zeitdimension

Visualisiert wird ein Netzwerk aus Publikationen (Knoten) und Zitationen (Kanten), wie es in der bibliografischen Analyse als Zitationsnetzwerk bekannt ist. Anstatt die Knoten mit einem simplen kraftgerichteten Layout anzuordnen, werden die Knoten in dieser speziellen Abwandlung des Zitationsgraphen entlang der dargestellten Zeitachse fixiert. Die Einordnung der Publikationen bezüglich dieses Constraints geschieht nach dem Jahr der Veröffentlichung. Die Idee für diese Abwandlung stammt aus der Arbeit *CitNetExplorer: A new software tool for analyzing and visualizing citation networks* (van Eck und Waltman, 2014) und dem daraus entstandenen Desktop-Programm CitNetExplorer, welches im Kapitel 3 vorgestellt wurde. Die Motivation zur Implementierung sind der neuartige Ansatz eines Constraints, die daraus folgende gute Übersichtlichkeit des Netzwerks und die bisher fehlende Portierung einer solchen Lösung für den Einsatz in Webbrowsern.

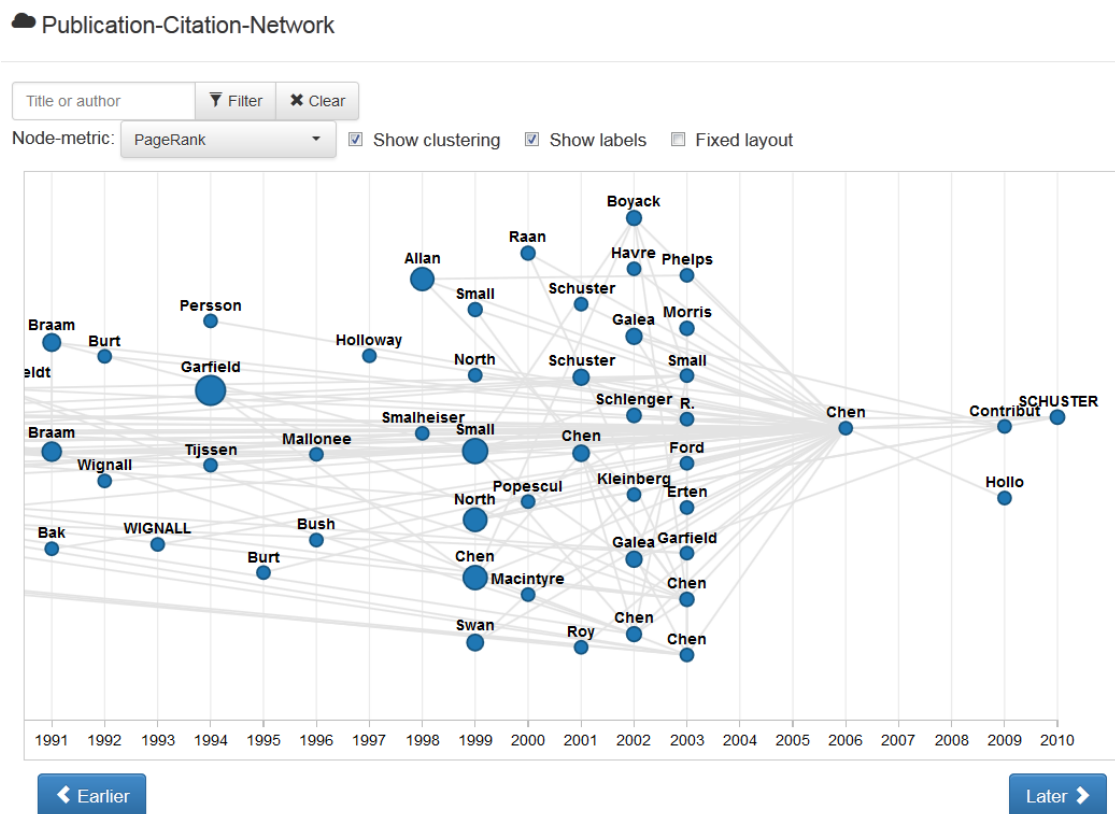


Abbildung 5.3.: Prototyp eines Zitationsnetzwerks mit Zeitdimension
Quelle: eigenes Werk (Prototyp, Screenshot)

Es bestehen folgende Möglichkeiten der Interaktion mit der Visualisierung:

1. Dargestellt ist immer ein Ausschnitt des vorliegenden Zeitbereichs. Ist dieser Zeitbereich zu groß, kann der gezeigte Ausschnitt durch ein Bewegen der Zeitachse mit der Maus oder durch die Schaltflächen „Earlier“ und „Later“ verschoben werden.

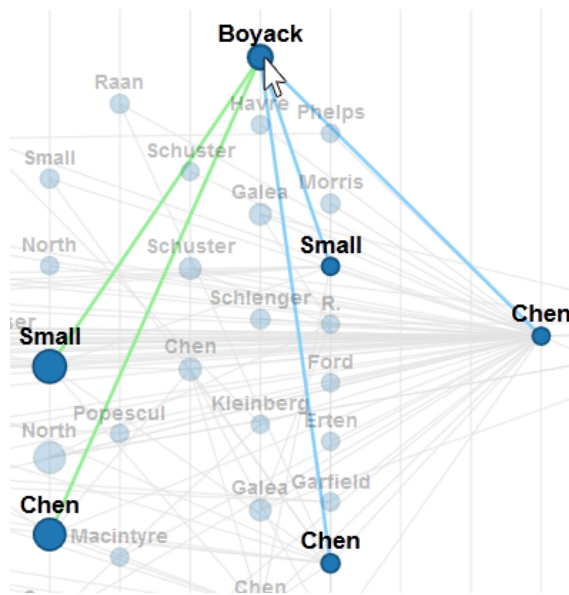


Abbildung 5.4.: Zitationsnetzwerk: Mausinteraktion mit einem Knoten. Die Publikation *Boyack* sowie alle ihre zitierten (grüne Kante) und zitierenden (blaue Kante) Publikationen werden hervorgehoben.
Quelle: eigenes Werk (Prototyp, Screenshot)

2. Mausinteraktionen mit einer Publikation (Knoten) heben diese Publikation und ihre zitierenden bzw. zitierten Publikationen hervor. Auch die entsprechenden Kanten werden hervorgehoben. Grüne Kanten stehen für eine zitierende Beziehung, blaue für eine zitierte Beziehung, jeweils bezogen auf die „anvisierte“ Publikation. Das restliche Netzwerk wird währenddessen teilweise ausgeblendet, um den Fokus zu verbessern. Das Ergebnis einer solchen Mausinteraktion ist der Abbildung 5.4 zu entnehmen.
3. Die Darstellung der verschiedenen farblich getrennten Cluster von Publikationen kann an- und abgeschaltet werden. Die Einteilung der Cluster wird durch *Edge-Betweenness-Clustering* vorgenommen. Die genaue Implementierung ist dem Kapitel 6 zu entnehmen.
4. Die Anzeige der Knotenbeschriftungen kann an- und abgeschaltet werden. Dies soll die Übersichtlichkeit verbessern.
5. Das Graph-Layout der Knoten ist wie beschrieben entlang der Zeitachse fixiert. Die vertikale Anordnung wird durch einen kraftgerichteten Algorithmus vorgenommen und wird im Browser berechnet. Die Knoten lassen sich mit der Maus bewegen.
6. Das kraftgerichtete Layout kann temporär an- und abgeschaltet werden. Dies ermöglicht es dem Benutzer, die Knoten auf Wunsch selbst anzuordnen und diese Anordnung beizubehalten.
7. Die Größe der Knoten kann durch verschiedene Metriken beeinflusst werden. Folgende Metriken sind neben einer Einheitsgröße geplant: *PageRank*, *Hub- und Authority-Score (HITS)*, *Eigenvektor Zentralität*, *Betweenness Zentralität* und *Closeness Zen-*

tralität. Die Folgen für die Netzwerkstruktur ergeben sich entsprechend dem Einsatzzweck der jeweiligen Metriken. Die Metriken werden zuvor bezogen auf das gesamte Netzwerk berechnet und können ohne Zeitverzögerung dargestellt werden. Weitere Einzelheiten zur Implementierung sind dem Kapitel 6 zu entnehmen.

8. Es steht auch ein Filtermechanismus zur Verfügung, der das Gesamtnetzwerk entsprechend der Eingabe reduziert. Anschließend wird der gefilterte Datensatz dargestellt. Unabhängig von Groß- und Kleinschreibung werden genau die Publikationen als Knoten gezeigt, die den Filterbegriff im Titel oder als Autor enthalten. Ein gesetzter Filter lässt sich ändern und entfernen.

5.1.3. Community Analyse

Eine Community Analyse ist immer dann sinnvoll, wenn der Begriff der Community oder der Cluster für eine Menge von Einheiten sinnvoll ist. Im vorliegenden Fall bezieht sich die Community Analyse insbesondere auf die Zusammenarbeit zwischen Autoren, auch Co-Autorschaft genannt. Zwei Autoren besitzen eine Co-Autorschafts-Beziehung, wenn diese zusammen an mindestens einer Publikation mitgewirkt haben. Dabei handelt es sich um Communities im klassischen Sinne, also der Verbindung von Personen durch gleiche Attribute. Diese systematischen Eigenschaften der Daten offenbaren eine Netzwerkstruktur. Dabei sind die Autoren als Knoten und deren Beziehungen als Kanten zu übersetzen. Die Communities entsprechen Clustern von Knoten des Netzwerks. Ein solches beispielhaftes Autoren-Netzwerk ist dem Entwurf der Visualisierung in der Abbildung 5.5 zu entnehmen. Dargestellt sind zehn fiktive Autoren und deren Co-Autorschafts-Beziehungen in drei farblich gekennzeichneten Communities.

Folgende Informationen lassen sich allgemein aus einem solchen Autoren-Netzwerk gewinnen:

1. Wo sind die Autoren lokalisiert und wie sind sie in Communities unterteilt?
2. Wie stark sind die Autoren innerhalb der Communities vernetzt und welche Autoren sind stark oder schwach vernetzt?
3. Wie stark ist die Vernetzung zwischen einzelnen Communities?
4. Gibt es autarke Communities, sprich isolierte Teilnetzwerke?
5. Welche Autoren leisten einen besonderen Beitrag an der Vernetzung zwischen den Communities?

Es lässt sich aus dieser Aufzählung folgern, dass sich die Art der gewonnen Erkenntnisse in verschiedene Teilbereiche gliedert: Erstens der Mikroebene (den Eigenschaften einzelner Autoren), zweitens der Mesoebene (den Eigenschaften der Communities als Aggregat der Autoren-Eigenschaften) und drittens der Makroebene (dem Verhältnis der Communities untereinander). Diese Vielfalt an Betrachtungsweisen ein und desselben Netzwerks erfordert eine geschickte Bereitstellung von Interaktionsmöglichkeiten mit der Visualisierung, denn ein starrer Fokus auf einen der Teilbereiche wäre den anderen beiden in ihrer Wirkung abträglich. Mit dieser Motivation wurden die folgenden Interaktionsmöglichkeiten als sinnvoll und umsetzbar kategorisiert:

1. Aufgrund der drei genannten Möglichkeiten der Betrachtung ist eine Zoom-Funktionalität implementiert, die eine Anpassung des Fokus zwischen den Teilbereichen ermöglicht. Auf diese Weise kann ein Betrachter, entsprechend seinen Anforderungen an die Visualisierung, selbstständig den Fokus verändern. Die Funktionalität ist durch die Schaltflächen rechts in der Abbildung 5.5 und durch das Scrollrad der Maus gegeben.
2. Zusammen mit einer Zoom-Funktion ist eine Navigation im Sichtbereich erforderlich, da es nun zu Fällen kommt, in denen sich nur Teile der Daten im Sichtbereich befinden. Diese Funktionalität ist durch eine „Drag-and-drop“ Interaktion mit dem Sichtbereich und durch die Schaltflächen rechts in der Abbildung gegeben.
3. Um die Usability zu verbessern, ist es möglich, durch die mittlere Schaltfläche in der Navigation rechts in der Abbildung den Ursprungszustand des Sichtbereichs wiederherzustellen.
4. Die Anordnung der Knoten und ihrer Communities geschieht automatisch durch einen kraft-gerichteten Algorithmus. Dieser kann deaktiviert werden, um die manuelle Anordnung der Knoten durch Mausinteraktion zu ermöglichen. Auf diese Weise können optisch verdeckte Teile des Netzwerks besser observiert werden.

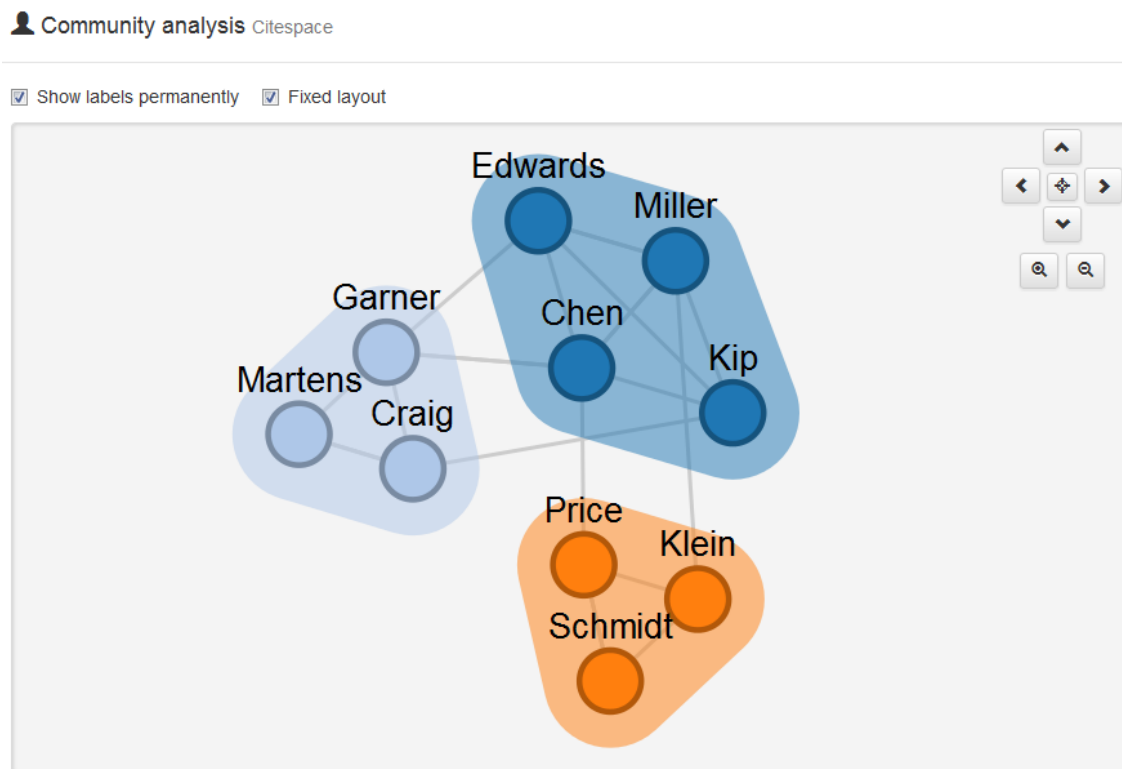


Abbildung 5.5.: Prototyp eines Community-Netzwerks. Dargestellt sind fiktive Co-Autorschafts-Beziehungen zwischen zehn Autoren.
Quelle: eigenes Werk (Prototyp, Screenshot)

5. Ebenfalls zur Verbesserung der Übersichtlichkeit ist es möglich, die dauerhafte Anzeige der Autorennamen auszuschalten. Stattdessen werden diese dann nur noch angezeigt, wenn der Autoren-Knoten mit der Maus überfahren wird.

5.1.4. Schlagwortwolke

Die Schlagwortwolke ist eine simple, aber effektive Art der Visualisierung, um einen ersten Überblick über eine große Menge von Informationen zu geben. In diesem Zusammenhang werden sie meist dazu verwendet, Gemeinsamkeiten von verteilten Informationen kompakt darzustellen. Aus diesem Grund kommen Schlagwortwolken auch in AcaVis zum Einsatz. Das Ziel ist dabei, den thematischen Tenor einer Sammlung von Publikationen übersichtlich darzustellen. Dazu können Informationen aus Titeln, Abstracts und zugeordneten Schlagworten in die Visualisierung einfließen. Ein prototypisches Beispiel einer solchen Schlagwortwolke in AcaVis ist der Abbildung 5.6 zu entnehmen.



Abbildung 5.6.: Schlagwortwolke als thematische Übersicht einer Sammlung von Publikationen.

Quelle: eigenes Werk (Prototyp, Screenshot)

Die Relevanz eines dargestellten Schlagwortes wird durch die beiden optischen Attribute der Schrift, Schriftgröße und Sättigung dargestellt. Die Anordnung ist zufällig, also mit keiner Semantik versehen. Um der Visualisierung weitere Aussagekraft zu verleihen, ist es mitunter notwendig, bestimmte Begriffe in der Betrachtung auszulassen. Dieses Problem ist jedoch sehr individuell bezüglich der zu Grunde liegenden Daten. Eine Interaktion des Benutzers mit der Visualisierung ist nicht vorgesehen.

5.2. Suchfunktion

Die Notwendigkeit einer Suchfunktion für ein Informationssystem zur bibliografischen Recherche ist ohne weitere Begründung offensichtlich. Nach dem eingangs gezeigten Dia-

gramm von Hull et al. (2008) werden in der hier beschriebenen Suchfunktion die Arbeitsschritte *Search*, *Select* und *Save (Metadata)* erfüllt. Für Informationssysteme mit Suchfunktion hat sich in den letzten Jahren eine gewisse Erwartungshaltung bezüglich der Funktionen eingestellt. Ursache ist die Allgegenwärtigkeit von webbasierten Suchmaschinen wie Google und Bing im täglichen Leben. Selbst konkurrierende Suchmaschinen stellen heute annähernd die gleichen Funktionen bereit. Prototypische Benutzeroberflächen der geplanten Suchfunktion und deren Ergebnisseite sind in den Abbildungen 5.7 und 5.8 zu sehen. Weitere Impressionen der Benutzeroberflächen sind im Anhang im Abschnitt A.1 zu finden.

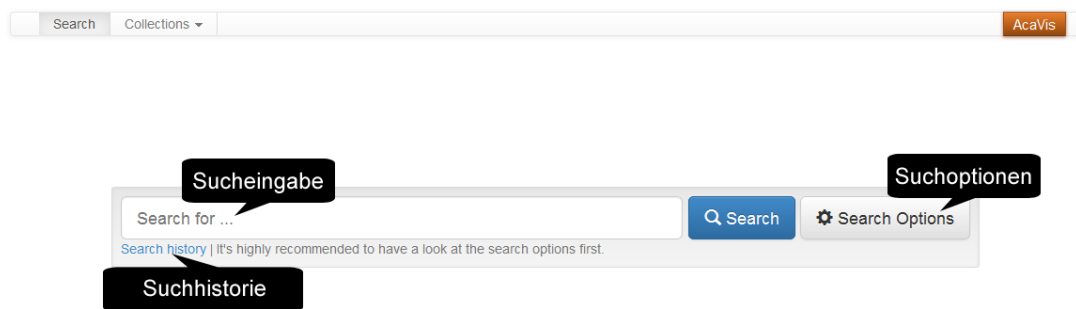


Abbildung 5.7.: Landingpage von AcaVis mit Sucheingabe

Primäres Ziel des Benutzers

Der Benutzer der Suchfunktion besitzt eine grobe Information. Er möchte diese Information nutzen, um relevante Publikationen zu erhalten. Er möchte zunächst sämtliche Publikationen angezeigt bekommen, die auf beliebige Art und Weise mit seiner Information verknüpft sind.

Der einfachste Bedienungsfall

Der Benutzer gibt seine Information in ein Suchfeld ein. Er muss keine weitere Einstellung vornehmen. Die Suche wird durch den Benutzer gestartet und liefert beliebig relevante Ergebnisse.

Erweiterungen des einfachsten Bedienungsfalls

1. Der Benutzer hat als spezielle Information den Namen eines Autors und möchte Publikationen dieses Autors angezeigt bekommen.
2. Der Benutzer möchte als Bedingung den Zeitraum eingrenzen, in dem die Publikationen erschienen sein müssen, um angezeigt zu werden.
3. Der Benutzer möchte die für ihn relevanten Publikationen einer Sammlung hinzufügen, um sie zu einem späteren Zeitpunkt genauer zu sichten.
4. Die Suchergebnisse möchte der Benutzer alphabetisch (Titel), nach Erscheinungsjahr und nach Relevanz sortieren.

Einstellungen

Die Erweiterungen des einfachsten Bedienungsfalls erfordern teilweise eine Einstellung seitens des Benutzers. Der Benutzer erwartet, dass die getätigten Einstellungen beim nächsten Aufruf der Seite noch immer vorhanden sind.

Unterstützende Visualisierungen

Bedingt durch die externe Schnittstelle zu MAS stehen an dieser Stelle keine detaillierten Informationen zu einer Publikation zur Verfügung. Lediglich der Titel, das Erscheinungsjahr, die Autoren, die Zahl der Zitationen und ggf. die Quelle sind verfügbar. Deshalb beschränkt sich die Visualisierung auf eine quantitative Darstellung der Verteilung der Publikationen im Suchergebnis nach Erscheinungsjahr, wie sie im Abschnitt 5.1.1 beschrieben wurde.

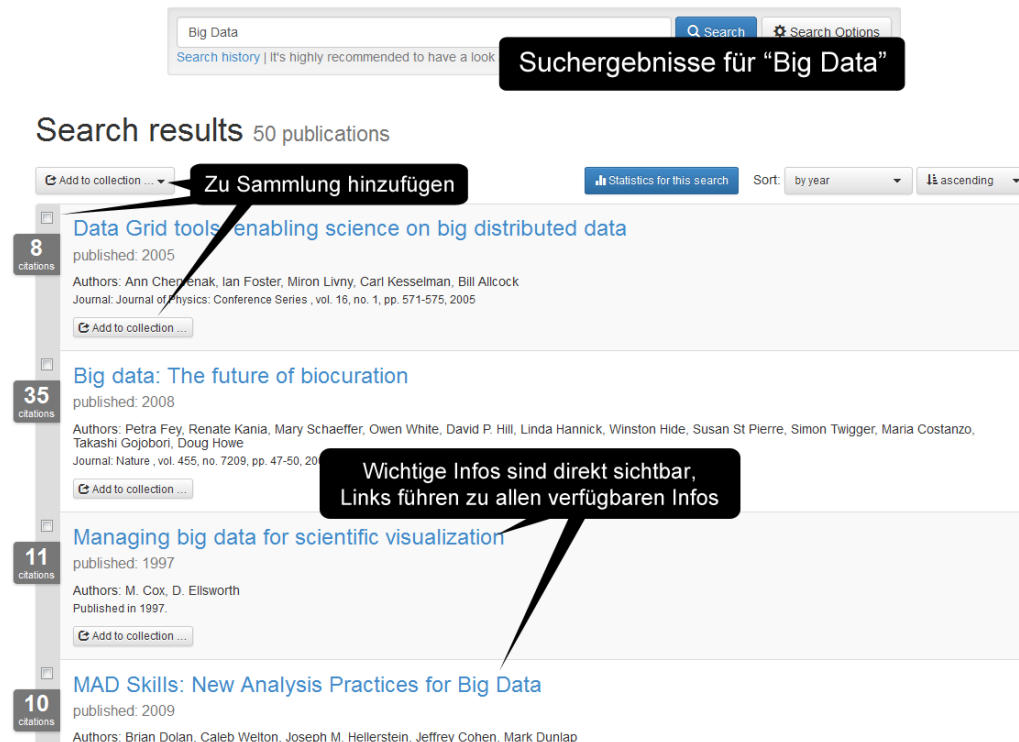


Abbildung 5.8.: Beispielhafte Ergebnisseite einer Suche nach „Big Data“ in AcaVis

5.3. Detailansicht

Die Notwendigkeit einer Detailansicht ergibt sich aus dem Bedürfnis bei der Recherche, eine Publikation genauer zu klassifizieren. Dieses Bedürfnis beschreiben auch Modjeska et al. (1996) in ihrem Paper *BIVTECI: A bibliographic visualization tool*, welches sich ebenfalls mit dem Aufbau eines bibliografischen Informationssystems unter Zuhilfenahme von Visualisierungen beschäftigt. Insbesondere werden dort theoretische Überlegungen ausgeführt. Ebenfalls sei laut den Autoren in diesem Zusammenhang wichtig, dass der Einfluss einer Publikation deutlich werde und diese Publikation in eine Chronologie eingeordnet werde.

Primäres Ziel des Benutzers

Der Benutzer ist im Besitz des genauen Identifikators⁶⁶ einer Publikation und möchte sämtliche verfügbaren Daten betrachten.

Der einfachste Bedienungsfall

Der Benutzer klickt auf einen Link zur vollständigen Darstellung aller verfügbaren Informationen einer Publikation. Die Darstellung wird angezeigt.

Erweiterungen des einfachsten Bedienungsfalls

1. Der Benutzer möchte alle zitierten oder zitierenden Publikationen einer Publikation als Auflistung betrachten.
2. Der Benutzer möchte wiederum die Detailansicht dieser Publikationen aus der Auflistung betrachten.
3. Der Benutzer möchte den Volltext der Publikation betrachten, direkt oder über eine verlinkte externe Quelle.
4. Da die Publikation noch nicht zwingend in einer Sammlung enthalten sein muss, möchte der Benutzer diese einer Sammlung hinzufügen können.

Einstellungen

Keine Einstellungen notwendig

Unterstützende Visualisierungen

Da an dieser Stelle sämtliche Informationen zu einer Publikation, inklusive zitierten und zitierenden Publikationen, über die MAS API verfügbar sind, kann ein Zitationsnetzwerk bezüglich dieser Menge von Publikationen erstellt werden. Die Umsetzung erfolgt entsprechend der Beschreibung des Abschnitts 5.1.2: *Zitationsnetzwerk mit Zeitdimension*. Die Merkmale dieser Visualisierung decken sich in etwa auch mit den vorgeschlagenen Eigenschaften einer Visualisierung für diesen Zweck, wie sie eingangs beschrieben wurden.

5.4. Sammlungen von Publikationen

Die Stärke von AcaVis soll darin liegen, mit Hilfe von Visualisierungen einem Benutzer gezielt strukturelle Aussagen über eine Menge von Publikationen zu liefern, welche darüber hinaus intuitiv verständlich sein sollen. Damit soll AcaVis die Teile einer Recherche automatisieren, die Fachkenntnisse der bibliografischen Visualisierung erfordern. Im Gegenzug kann AcaVis dem Benutzer eine Aufgabe jedoch noch nicht abnehmen: Die Auswahl der (vermeintlich) relevanten Publikationen. Die existierenden groben Kategorien nach Disziplinen, Journalen, Konferenzen, usw. sind für diesen Zweck nicht hinreichend. Eine handverlesene Auswahl, wie sie auch Modjeska et al. (1996) vorschlagen, ist hier notwendig. Deshalb soll es AcaVis ermöglichen, individuelle Sammlungen von Publikationen anzulegen und diese dauerhaft zu speichern. Eine prototypische Benutzeroberfläche der geplanten Sammlungen ist in der Abbildung 5.9 zu sehen. Weitere Impressionen der Benutzeroberflächen sind im Anhang im Abschnitt A.2 zu finden.

⁶⁶Ein Identifikator kann zum Beispiel ein beliebiger Hyperlink zu den vollständigen Informationen genau einer Publikation sein.

The screenshot displays the Citespace web application interface. At the top, there's a search bar labeled 'Temporärer Filter' and a title 'Übersicht der Sammlung "Citespace"'. Below the search bar, there are filters and sorting options. The main area lists three publications:

- CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature (2005)** by Chaomei Chen. It has 121 citations and is associated with disciplines like Computer Science, Data Mining, and Information Retrieval.
- CiteSpace II: Visualization and Knowledge Discovery in Bibliographic Databases (2005)** by Marie B. Synnæstvedt, Chaomei Chen, and John H. Holmes. It has 5 citations and is associated with Computer Science and Information Retrieval.
- Visualization Analysis of Dynamic Evolution in the Field of National Innovation Studies (2010)** by Liu Fengchao, Xu Qian, and Xia Xinchun. It has 0 citations.

On the right side, there's a sidebar with 'Tag-cloud Top 20' and 'Overview' statistics. The 'Overview' section shows: Total 165 publications, 66 authors, Timespan 1973 – 2013. Under 'Visualizations', there are buttons for 'Timeline of publications', 'Publication-Citation-Network', 'Co-Keyword Analysis', and 'Community Analysis'.

Abbildung 5.9.: Beispielhafte Sammlung von Publikationen in AcaVis. Die verfügbaren Visualisierungen wurden im Abschnitt 5.1 ausführlich beschrieben.

Primäres Ziel des Benutzers

Der Benutzer hat eine Sammlung in einem früheren Arbeitsschritt angelegt, um die Publikationen später genauer zu betrachten. Er möchte die Publikationen nun in der Detailansicht wie in der Übersicht betrachten, um daraus Informationen zu gewinnen und die vermeintlich relevanten Publikationen auf die wirklich relevanten Publikationen reduzieren.

Der einfachste Bedienungsfall

Der Benutzer öffnet die Sammlung und betrachtet die Übersicht. Bei weiterem Interesse öffnet der Benutzer die zugehörige Detailansicht einer Publikation. Ist diese Publikation nicht relevant, möchte sie der Benutzer aus der Sammlung löschen.

Erweiterungen des einfachsten Bedienungsfalls

1. Der Benutzer möchte eine neue Sammlung anlegen oder eine bestehende entfernen können.
2. Der Benutzer möchte Publikationen in andere Sammlungen verschieben oder kopieren können.
3. Der Benutzer möchte Publikationen aus einer Sammlung löschen können.
4. Der Benutzer möchte die Sortierung der Publikationen in der Sammlung alphabetisch (Titel), nach Erscheinungsjahr oder nach Relevanz (durch Zitationen) vornehmen können.

5. Der Benutzer möchte, falls möglich, mehr Details zur Publikation sehen als in den Suchergebnissen. Die Darstellung soll aber noch übersichtlich bleiben und nicht die Detailansicht ersetzen.

Einstellungen

Keine Einstellungen notwendig

Unterstützende Visualisierungen

Da es sich bei der Erstellung von Sammlungen um die Kernfunktion von AcaVis handelt, werden hier insgesamt vier verschiedene Visualisierungen bereitgestellt. Erstens eine Übersicht über die Verteilung der Publikationen in der Sammlung über die Zeit, wie sie im Abschnitt 5.1.1 beschrieben wurde. Zweitens ein Zitationsnetzwerk mit Zeitdimension nach Abschnitt 5.1.2. Drittens eine Visualisierung zur Analyse von Autoren-Communities bezüglich der Sammlung von Publikationen, wobei die Co-Autorschaft über die Beschaffenheit der Communities entscheidet. Diese Visualisierung wurde im Detail im Abschnitt 5.1.3 beschrieben. Viertens und letztens ist direkt auf der Übersichtsseite einer Sammlung eine Schlagwortwolke enthalten. Diese visualisiert die gemeinsamen Schlagworte aller Publikationen einer Sammlung. Die Schlagwortwolke wurde im Abschnitt 5.1.4 betrachtet.

5.5. Anwendungsfälle mit bedingter Umsetzbarkeit und Erweiterungen

Wie bereits ausführlich begründet, ist der behandelte Themenbereich sehr umfangreich. Neben den dargestellten Anwendungsfällen, die eher eine notwendige Grundfunktionalität bereitstellen, gibt es noch weitere Anwendungsfälle sowie Erweiterungen dieser Anwendungsfälle in vielerlei Ausprägung. Da die Möglichkeiten der Umsetzung innerhalb dieser Arbeit begrenzt sind, können diese in AcaVis noch nicht vollständig implementiert werden. Trotzdem sollen einige mögliche Erweiterungen skizziert werden, die in eine Weiterentwicklung von AcaVis einfließen könnten.

5.5.1. Erweiterung der Suchfunktion

Einige Vorschläge einer komplexen Suchfunktion für den Zweck der Recherche sind zusätzlich zu den Visualisierungen auch in den Skizzen des ALADIN-Projekts (Humboldt Institut für Internet und Gesellschaft – ohne Autor, 2014) enthalten, welches auch eine konzeptionelle Basis von AcaVis liefert. Neben der Suche innerhalb eines Dokuments und seiner Metadaten wird konkret vorgeschlagen, dass die zu durchsuchenden Datenfelder, neben der Standard-Suche bezüglich aller Datenfelder, explizit ausgewählt werden können (Humboldt Institut für Internet und Gesellschaft – ohne Autor, 2014, Folie 5). Außerdem sollten Möglichkeiten geschaffen werden, eine Menge von Suchergebnissen entsprechend weiterer Attribute zu filtern. Insbesondere werden hier die Einschränkung auf eine Auswahl von wissenschaftlichen Disziplinen, auf den Typ der Veröffentlichung einer Publikation (Buch, Journal, Paper, usw.) und auf eine Auswahl von Autoren genannt (Humboldt Institut für

Internet und Gesellschaft – ohne Autor, 2014, Folie 5). Diese Erweiterungen sind mit angemessenem Aufwand mit AcaVis umsetzbar. Auch die notwendigen Informationen und Schnittstellen stünden in MAS prinzipiell bereit.

Zudem wäre die Visualisierung eines Suchergebnisses durch eine Schlagwortwolke hilfreich. Darin ließen sich wichtige Begriffe des Themenfelds wiederfinden, die dem Benutzer bis dato noch nicht geläufig sind. So kann der Benutzer im Anschluss an eine eher allgemein formulierte Suchanfrage eine speziellere Suchanfrage stellen, die durch die Verwendung von Fachbegriffen im Vergleich mehr Ergebnisse oder Ergebnisse mit größerer Relevanz liefert. Leider ist eine solche Funktion mit MAS als Datenbasis nur mit relativ viel Aufwand umsetzbar. Die hierfür nötigen Daten sind jedoch prinzipiell verfügbar.

Es wäre bei der Betrachtung von Ergebnissen einer Suche auch hilfreich, wenn zu einem Suchergebnis angezeigt wird, ob diese Publikation bereits in einer Sammlung enthalten ist und falls ja, in welcher oder welchen Sammlungen. Die Umsetzbarkeit ist nicht an eine externe Datenbasis gebunden. Somit kann diese Erweiterung mit geringem Aufwand direkt in AcaVis implementiert werden.

5.5.2. Erweiterung der Detailansicht

Die Detailansicht würde deutlich bereichert, wenn dort auch die Volltexte der Publikationen zur Verfügung stünden. Diese Anforderung ist leider aufgrund der Verfügbarkeit dieser Daten und den besprochenen Interessens- und Lizenzkonflikten auf absehbare Zeit nicht umsetzbar. Als Alternative könnten Verlinkungen zu den Dokumenten in kostenlosen oder kostenpflichtigen Verzeichnissen dienen. MAS stellt solche Verlinkungen bereit, aber nur wenn die zugehörige Publikation kostenlos und öffentlich verfügbar ist.

Weiterhin kann die Detailansicht durch verschiedene Visualisierungen rund um die jeweilige Publikation erweitert werden. Anregungen und Beispiele sind zum Beispiel im Kapitel 3 dieser Arbeit zu finden. Darüber hinaus existiert eine Vielzahl von weiteren theoretischen und praktischen Ansätzen.

5.5.3. Erweiterung der Sammlungen von Publikationen

Zur Organisation großer Sammlungen kann es nützlich sein, diese temporär zu gruppieren oder zu filtern. Ein Filtermechanismus sollte in der Lage sein, die enthaltenen Publikationen entsprechend der wichtigsten Datenfelder wie Titel, Autoren, wissenschaftliche Disziplin, Schlagworte oder der Relevanz zu filtern. Eine solche Filterung soll sich ohne Datenverlust aufheben lassen. Die Sammlung stellt zwar im übertragenen Sinn selbst eine Gruppe aller bestehenden Publikationen dar. Es kann jedoch sinnvoll sein, eine Sammlung in verschiedene weitere Gruppen zu unterteilen, ohne für jede Gruppe eine eigene Sammlung erstellen zu müssen. Dieses Merkmal würde die Organisation großer Sammlungen erleichtern. Die Gruppenzugehörigkeit sollte „leichtgewichtig“ über ein kategorisierendes optisches Attribut dargestellt werden, also zum Beispiel durch eine Markierung in variierendem Farbton. Die Umsetzung ist mit angemessenem Aufwand in AcaVis realisierbar. Es werden keine externen Datenquellen benötigt.

Um den individuellen Charakter einer Sammlung zu stärken, sollte es dem Benutzer möglich sein, eine Sammlung oder einzelne enthaltene Publikationen mit Anmerkungen zu versehen. Diese können bei einer nachfolgenden Sichtung oder Einordnung behilflich sein. Als erweitertes Element könnten diese Anmerkungen wiederum analysiert und visualisiert werden, sodass eine automatisierte Zusammenfassung auf der Basis der individuellen Anmerkungen eines Benutzers erstellt werden kann. Eine solche Verfügbarkeit von Anmerkungen ist in AcaVis mit angemessenem Aufwand umsetzbar.

Um den Vorgang der Erstellung von Sammlungen zu vereinfachen, können Funktionen implementiert werden, welche Publikationen nach einem vorgegebenen Muster automatisch hinzufügen. Das könnten zum Beispiel alle zitierten Publikationen einer Publikation sein. Optional könnte es hilfreich sein, nur solche Publikationen hinzuzufügen, die einem bestimmten Gütekriterium genügen. Solche Kriterien könnten zum Beispiel das Erscheinungsjahr, die Forschungsdisziplin oder die Anzahl der Zitationen einer Publikation sein. Je nach Umfang der beschriebenen Erweiterung kann diese in der Implementierung durchaus auch mit großem Aufwand verbunden sein, je nach Ausprägung des gewünschten Daten-Minings.

5.5.4. Übergreifende Erweiterungen

AcaVis wird im Rahmen dieser Arbeit prototypisch erstellt. Daraus folgt erstens, dass an vielen Stellen Potenzial zur Optimierung der Implementierung besteht. Zweitens wird aus diesem Grund zunächst auf eine Benutzer- und Rechteverwaltung verzichtet. Um AcaVis öffentlich beziehungsweise mit mehreren Benutzern gleichzeitig betreiben zu können, wäre es notwendig, eine Benutzer- und Rechteverwaltung einzuführen. Dadurch können die Sammlungen eines Benutzers gegen Zugriff und Veränderung geschützt werden. Eine Benutzer- und Rechteverwaltung ist aufgrund der vorgesehenen Struktur von AcaVis und unter Rückgriff entsprechender Bibliotheken mit angemessenem Aufwand realisierbar.

6. Die Implementierung von AcaVis

Nachdem im vorangegangenen Kapitel die Anwendungsfälle eingehend erläutert wurden, wird nun die daraus entstandene Implementierung im Detail erklärt. Anwendungsfälle schildern lediglich die Eingaben und deren Ergebnisse auf einer abstrakten Ebene und ohne die Details der Umsetzung zu definieren. Ein lauffähiges Programm das diesen Anwendungsfällen Rechnung trägt, beinhaltet aber gerade eine konkrete Umsetzung. Zunächst wird ein grober Überblick über die Architektur von AcaVis gegeben. Weiter wird kurz auf die wichtigsten externen Softwarebibliotheken eingegangen, die in AcaVis verwendet wurden. Schließlich werden noch einige Hinweise gegeben, wie AcaVis gewartet und weiterentwickelt werden kann.

6.1. Programmarchitektur

AcaVis ist entsprechend dem Anforderungsprofil der Aufgabenstellung als webbasierte Anwendung umgesetzt. Als Programmiersprache für das Backend ist Java verwendet worden. Diese Designentscheidung ist unter anderem der weiten Verbreitung und der vielfältigen Möglichkeiten im Bereich der professionellen Anwendungen⁶⁷ geschuldet. Für die Entwicklung von Webanwendungen in Java existieren etablierte Konzepte und Frameworks, die eine Umsetzung erleichtern. Außerdem ist Java als Programmiersprache im universitären Umfeld verbreitet. So wird eine Fortentwicklung des Projekts bei Bedarf sichergestellt.

Aufgrund der vielen Unwägbarkeiten im Zusammenhang mit der Anbindung zu externen Informationssystemen – hier sind besonders die allgemeine schlechte Verfügbarkeit von Informationen und die Defizite der MAS Datenbank zu nennen – sollte die Architektur von AcaVis derartig ausgeführt werden, dass die Funktionalität entkoppelt von einem bestimmten Informationssystem ist. In AcaVis integrierte Anbindungen zu externen Informationssystemen sollen wartbar, austauschbar und selektiv sein. Als Selektivität wird hier im Folgenden die Auswahlmöglichkeit einer der verfügbaren integrierten Anbindungen zur Laufzeit bezeichnet. Das impliziert auch, dass es möglich sein soll, mehrere Anbindungen parallel in AcaVis zu betreiben. Es wird aber explizit darauf verzichtet, die Ergebnisse verschiedener Informationssysteme miteinander zu verknüpfen. Diese Aufgabe ist im Rahmen dieser Arbeit nicht zu leisten, ist aber mit den verfügbaren Daten prinzipiell möglich.

Um das beschriebene Verhalten in AcaVis zu erreichen, wird die Architektur nach dem typischen Drei-Schicht-Modell der Softwareentwicklung in die Schichten *Integration*, *Service* und *Presentation* aufgeteilt. Eine schematische Darstellung ist der Abbildung 6.1 zu

⁶⁷„Professionelle“ Anwendungen oder auch Enterprise-Anwendungen in Java werden durch die Java EE Spezifikation definiert. Diese beinhaltet insbesondere Konzepte für die Entwicklung von Webanwendungen und die Verknüpfung mit Datenbanken.

entnehmen. In der Integration-Schicht sind die direkten Anbindungen zu den externen Informationssystemen implementiert. Derzeit ist MAS als Informationssystem implementiert. Die Service-Schicht stellt die Daten der Integration-Schicht über definierte Schnittstellen bereit. Auch die Verwaltung von Sammlungen und Einstellungen eines Benutzers wird hier implementiert. Die Presentation-Schicht bedient sich dieser einheitlichen Daten aus der Service-Schicht und ist für deren Darstellung als Webapplikation zuständig. Auch zwischen der Integration-Schicht und der Service-Schicht sind entsprechende Schnittstellen definiert. Diese stellen sicher, dass Anbindungen zu externen Informationssystemen gleichartig genutzt werden können. Durch diese Konventionen wird außerdem auf beiden Seiten der Schnittstelle sichergestellt, dass fehlende Teil-Informationen die davon unabhängigen Funktionen nicht beeinträchtigen.

Die Schnittstelle zwischen Integration-Schicht und Service-Schicht definiert außerdem die Container-Klassen, die der einheitlichen Repräsentation von Publikationen dienen. Die Schnittstelle unterscheidet hier zwei Container-Klassen für Publikationen. Die erste trägt den Namen *PublicationSearchData* und repräsentiert eine Publikation als Suchergebnis, die zweite trägt den Namen *PublicationData* und repräsentiert eine vollständige Publikation mit sämtlichen verfügbaren Informationen. Außerdem existiert eine Container-Klasse mit dem Namen *PublicationIdentifier*, welche in der Lage ist, verschiedene Identifikatoren gleichzeitig aufzunehmen. Folgende Datenfelder sind derzeit in den jeweiligen Container-Klassen enthalten:

PublicationSearchData

Identifikator, Titel, Erscheinungsjahr, Autoren (ungeordnete Liste), Metrik, Name der Metrik (falls vorhanden), Text-Vorschau und Quelle

Anmerkung: Bis auf den Identifikator und den Titel sind alle Felder optional

PublicationData

Identifikator, Titel, Erscheinungsjahr, Autoren (geordnete Liste), Abstract, Quelle, Schlagworte (Liste), URLs zum Volltext (Liste), Disziplin, Unterdisziplin, zitierte Publikationen (Liste), zitierende Publikationen (Liste), Anzahl der zitierten Publikationen, Anzahl der zitierenden Publikationen

Anmerkung: Bis auf den Identifikator und den Titel sind alle Felder optional

PublicationIdentifier

MAS-Primärschlüssel, Mendeley-Primärschlüssel, DOI-Identifikator, ISBN-Nummer

Anmerkung: Alle Felder sind optional. Wird ein Identifikator abgefragt, ist aber nicht verfügbar, wird eine passende Exception erstellt.

Um Anfragen zu beschleunigen und eine Persistenz der Informationen zu ermöglichen, wird in AcaVis auf eine Datenbank zurückgegriffen. Die derzeit verwendete Datenbank ist direkt in die Programmpakete integriert und erfordert keine separate Installation oder Konfiguration. In dieser Datenbank werden Suchergebnisse zeitlich begrenzt zwischengespeichert (Caching) und die Informationen der erstellten Sammlungen dauerhaft gespeichert (Persistenz). Die bereitgestellte Datenbank lässt sich leicht durch eine andere Datenbank ersetzen. Für einen dauerhaften Produktiveinsatz von AcaVis wird ohnehin eine externe Datenbank empfohlen.

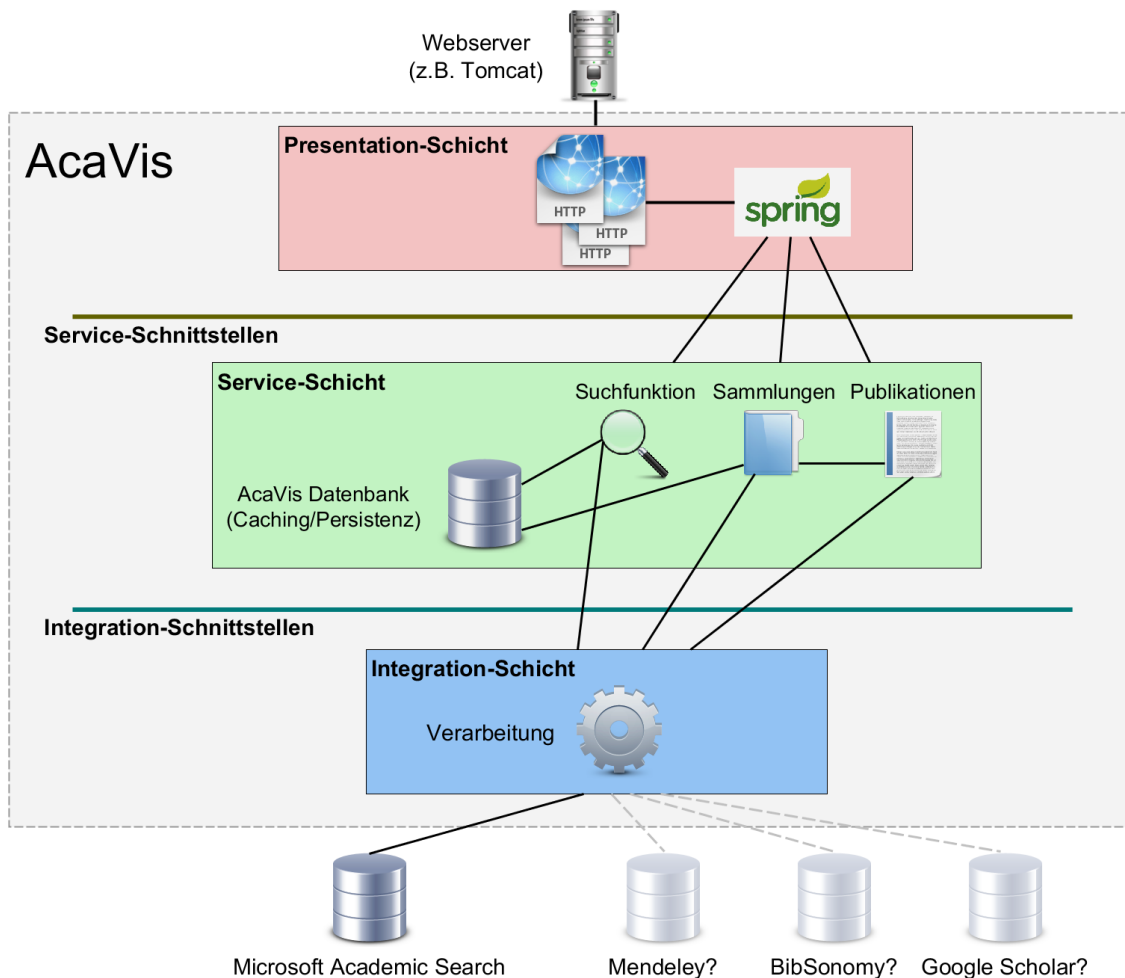


Abbildung 6.1.: Darstellung der Drei-Schicht-Architektur von AcaVis, den externen Anbindungen und der wichtigsten Komponenten. Die Informationssysteme Mendeley, BibSonomy und Google Scholar stehen hier nur exemplarisch für weitere mögliche Implementierungen.
Quelle: eigenes Werk

AcaVis setzt in der Präsentation-Schicht auf die Bibliothek Spring Framework, genauer den Bestandteil Spring MVC für Webanwendungen. Diese Bibliothek ermöglicht eine Aufteilung der Ressourcen in Models, Views und Controller, entsprechend dem klassischen MVC-Entwurfsmuster. Die Besonderheit ist hier, dass es sich bei den Views nicht um Java-Klassen sondern um Templates handelt. In diesem Zusammenhang wird auf Templates im Format JSP (JavaServer Pages) zurückgegriffen. Zusammen mit der Skriptsprache JSTL lassen sich so mit JSPs komplexe dynamische Webseite erzeugen. Spring übernimmt außerdem weite Teile der Verwaltung von Ressourcen, Datenbankverbindungen, Cookies und Fehlerbehandlung. Es stellt außerdem umfangreiche Werkzeuge zur individuellen Gestaltung zur Verfügung.

6.2. Externe Bibliotheken

Wie in der Softwareentwicklung üblich, wurden in AcaVis diverse externe Bibliotheken verwendet, um die Entwicklung zu beschleunigen. Um eine Verwaltung der Bibliotheken besonders einfach zu gestalten, werden diese mit Apache Maven eingebunden. Die verwendeten Bibliotheken sind dementsprechend in der Datei `pom.xml` zu finden. Alle verwendeten Bibliotheken sind kostenlos verfügbar. Die wichtigsten sollen hier kurz genannt werden, um Anhaltspunkte für die Weiterentwicklung zu liefern. Zunächst werden jene Bibliotheken aufgeführt, die im Backend in der Java-Programmierung verwendet wurden:

Spring Framework

Das Spring Framework⁶⁸ (kurz Spring) ist ein mächtiges Hilfsmittel bei der Umsetzung der Konzepte aus der Java EE Spezifikation. Besonders die gute Skalierbarkeit war bei der Wahl ein ausschlaggebender Faktor. Spring ist sehr umfangreich, was viele Programmierer kleiner Projekte von der Verwendung abschreckt. In AcaVis wurden aber lediglich die gängigsten Funktionen der Webentwicklung verwendet und keines der komplexeren „Enterprise-Konzepte“.

Verwendete Version: Spring Framework 3.1.1

HyperSQL Database

Die HyperSQL Database⁶⁹ (kurz HSQLDB) ist eine Datenbanksoftware, die in Java geschrieben ist. HSQLDB wurde in AcaVis verwendet, weil die Datenbank auf Dateisystembasis direkt in eine Java-Applikation integriert werden kann. Es ist also zunächst keine externe Datenbank für den Betrieb notwendig. Außerdem bietet HSQLDB eine Unterstützung für sogenannte *In-Memory Datenbanken*. Diese bestehen während der Laufzeit im Arbeitsspeicher und werden bei Programmende gelöscht oder im Dateisystem zwischengespeichert. Eine HSQLDB In-Memory Datenbank kommt bei der lokalen Nachbildung der Join-Funktionalität für MAS in AcaVis zum Einsatz.

Verwendete Version: HSQLDB 2.3.2

Hibernate/JPA

Hibernate⁷⁰ ist eine Provider-Bibliothek für Datenbankverbindungen in Java. Es implementiert die Spezifikation der Java Persistence API (kurz JPA). Diese Implementierung kommt zusammen mit den zugehörigen JPA-Annotationen in AcaVis für sämtliche Datenbankoperationen zum Einsatz.

Verwendete Version: Hibernate 4.0, JPA 2 Support

OData4J

OData4J⁷¹ ist eine Java-Bibliothek für den Umgang mit online verfügbaren Datenbanksystemen, die den Zugriff über eine OData-Schnittstelle erlauben. Die MAS API verwendet ausschließlich eine OData-Schnittstelle für den Transfer der Datensätze, was die Verwendung einer entsprechenden Bibliothek erforderlich macht.

Verwendete Version: OData4J 0.7

⁶⁸Spring Framework <http://projects.spring.io/spring-framework/>

⁶⁹HSQLDB <http://hsqldb.org/>

⁷⁰Hibernate <http://hibernate.org/orm/>

⁷¹OData4J <http://odata4j.org/>

jsoup

Die jsoup-Bibliothek⁷² ermöglicht das Auslesen von Informationen aus XML-strukturierten Dokumenten, insbesondere von HTML-Dokumenten, mittels einer Selektor-basierten Ausdruckssprache. Die Ausdruckssprache ist nahezu identisch mit jener auf denen die CSS-Selektoren in Cascading-Stylesheets basieren. Die Bibliothek wird für den Webcrawler der regulären MAS-Suchmaschine verwendet.

Verwendete Version: jsoup 1.7.3

JUNG Framework (Java Universal Network/Graph Framework)

Das JUNG Framework⁷³ stellt eine Bibliothek für die Modellierung und Analyse von Graphen bereit. Da viele der in AcaVis integrierten Visualisierungen einen Netzwerkcharakter besitzen, liegt es nahe diese als Graphen zu modellieren. Weiter stellt das JUNG Framework verschiedene gängige Algorithmen zur Berechnung von Netzwerkmetriken bereit. Die Implementierungen der folgenden Algorithmen in AcaVis verwenden direkt die jeweilige Implementierung des JUNG Frameworks: Page-Rank, Hubs and Authorities (HITS), Betweenness-Zentralität, Closeness-Zentralität, Eigenvektor-Zentralität und Edge-Betweenness Clustering. Anmerkungen zu diesen Algorithmen sowie zu den entsprechenden Publikationen sind der Dokumentation des JUNG Frameworks⁷⁴ zu entnehmen.

Verwendete Version: JUNG 2.0.1

Ergänzend dazu folgt eine Übersicht der Bibliotheken die im Frontend, also innerhalb der Webdarstellung, verwendet wurden:

jQuery

Die jQuery-Bibliothek⁷⁵ ist ein nützliches Hilfsmittel bei der Programmierung mit Javascript in Webanwendungen. Es stellt eine Vielzahl von Hilfsmethoden bereit, die häufig auftretende Probleme in der Javascript-Programmierung lösen. Von besonderem Wert ist jQuery auch, weil es die deutlich variierenden Verhaltensweisen der gängigen Webbrowser angleicht und so eine für den Programmierer verlässliche Funktionssammlung darstellt. Eigene Workarounds entfallen dadurch fast vollständig. Außerdem wird jQuery von den beiden hier ebenfalls verwendeten Bibliotheken Twitter Bootstrap und D3.js benötigt.

Verwendete Version: jQuery 1.11.1

Twitter Bootstrap

Twitter Bootstrap⁷⁶ (oder auch nur Bootstrap) ist eine Bibliothek zur Entwicklung von Benutzeroberflächen in Webanwendungen. Bootstrap stellt für sämtliche grafischen Elemente von Webseiten einen angepassten und optisch aufgeräumten Standardstil bereit. Diese Aufgaben entfallen bei der Verwendung für den Webentwickler. Bei Bedarf können Individualisierungen vorgenommen werden. Bootstrap beschleunigt insbesondere die Erstellung von ansprechenden und aufgeräumten Prototypen,

⁷²jsoup <http://jsoup.org/>

⁷³JUNG Framework <http://jung.sourceforge.net/index.html>

⁷⁴JUNG Framework Dokumentation <http://jung.sourceforge.net/doc/api/index.html>

⁷⁵jQuery <http://jquery.com/>

⁷⁶Twitter Bootstrap <http://getbootstrap.com/>

was für die Verwendung in AcaVis ausschlaggebend war.

Verwendete Version: Bootstrap 3.2.0

D3 (Data-Driven Documents)

Die D3-Bibliothek⁷⁷ stellt mächtige Mechanismen zur Erstellung von Visualisierungen in Webanwendungen bereit. Diese werden mit Hilfe von Javascript und den mittlerweile in vielen Webbrowsern verfügbaren SVG-Elementen realisiert. Quasi sämtliche Visualisierungen sind mit D3 realisierbar. Eine Vielzahl von Beispielen ist auf der Projektseite zu finden. Neben klassischen „starren“ Visualisierungen ermöglicht D3 die Erstellung von interaktiven, oder sich verändernden Visualisierungen. Die große Anzahl von Möglichkeiten und Beispielen waren ausschlaggebend für die Verwendung der D3-Bibliothek in AcaVis.

Verwendete Version: D3.js 3.4.11

Die verwendeten Bibliotheken erlauben es, mit AcaVis trotz des prototypischen Anspruchs eine professionelle und auf anerkannten Techniken und Algorithmen basierende Implementierung zu schaffen. Ohne die Verwendung der Bibliotheken hätte deutlich weniger Entwicklungsarbeit in die Kernziele von AcaVis fließen können.

6.3. Java-Pakete und wichtige Verzeichnisse

Ergänzend zur beschriebenen Programmarchitektur sind in diesem Abschnitt wichtige Java-Pakete und Verzeichnisse mit deren Verwendungszweck aufgelistet. Im nachfolgenden Abschnitt 6.4 wird teilweise auf diese Pakete verwiesen.

Integration-Schicht

`de.kbs.acavis.integration`

Schnittstellen zur Service-Schicht und entsprechende Exceptions.

`de.kbs.acavis.integration.controller`

Enthält direkte Implementierungen der Schnittstellen pro externem Informationssystem.

`de.kbs.acavis.integration.mas.*`

Implementierung der Anbindung an die MAS API und Webcrawler.

`de.kbs.acavis.integration.model`

Die Container-Klassen des Datentransfers für die Schnittstellen zur Service-Schicht.

Service-Schicht

`de.kbs.acavis.service`

Controller der Metaebene von AcaVis.

`de.kbs.acavis.service.controller`

Die Controller für die bereitgestellten Funktionalitäten in und durch AcaVis.

⁷⁷D3 (Data-Driven Documents) <http://d3js.org/>

`de.kbs.acavis.service.model.persistence`

Datenbank-Entitäten für die Speicherung von Sammlungen und Suchvorgängen.

`de.kbs.acavis.service.model.visualization`

Container für Netzwerk- und Diagrammdaten.

Presentation-Schicht

`de.kbs.acavis.presentation.controller`

Request-Controller des Web-Application-Routings (Spring).

`de.kbs.acavis.presentation.model`

Containerklassen für die Daten von Formular-Elementen im View.

`/src/main/webapp/resources`

Web-Ressourcen: Grafiken, Stylesheets, Javascript und Web-Bibliotheken

`/src/main/webapp/WEB-INF/views`

`/src/main/webapp/WEB-INF/tags`

Alle relevanten Daten der Views: JSP-Dateien und JSP-Tags.

`/src/main/webapp/WEB-INF/spring/appServlet`

Enthält die Hauptkonfiguration von Spring.

6.4. Wartung und Weiterentwicklung

Die Implementierung von AcaVis ist für ein derart kleines Projekt vergleichsweise komplex. Um eine etwaige Weiterentwicklung und Wartung des Systems zu ermöglichen, werden an dieser Stelle noch einige Elemente einer technischen Dokumentation ergänzt. Diese Aufzählung umfasst nur die wichtigsten und wahrscheinlichsten Fälle. Die Quelltexte sind jedoch kommentiert, sodass sich sämtliche funktionalen Eigenschaften von AcaVis, ändern lassen. Anzumerken ist, dass sich die folgenden Ausführungen teilweise direkt auf die Struktur der AcaVis Programm-Pakete und Klassen bezieht.

Der MAS Webcrawler

Die Suchfunktion in MAS basiert wie erwähnt auf einem Webcrawler, der die reguläre Suchfunktion auf der Webseite von MAS ansteuert und die Ergebnisse ausliefert. Es kann in Zukunft notwendig werden, dass entweder die URL des Dienstes oder die Selektoren die für das Auslesen der Daten zuständig sind, geändert werden müssen. Diese Eigenschaften lassen sich in der Klasse `WebRequests` des Pakets `de.kbs.acavis.integration.mas.controller` ändern. Die Felder sind als Konstanten ausgeführt und tragen entsprechende Kommentare zu deren Verwendung. Auch das Verbindungs-Timeout lässt sich hier anpassen. Derzeit ist ein Wert von 15 Sekunden eingestellt.

OData Anfragen an MAS

Die URL des MAS-OData-Dienstes lässt sich in der Klasse `MASOdataConnector` des Pakets `de.kbs.acavis.integration.mas.odata` konfigurieren. Ebenfalls in diesem Paket befinden sich Hilfsklassen für die Abfrage von Daten aus der MAS API. Alle Anfragen sollten immer mit den beiden Hilfsklassen `MASEntityHelper` und `MASRelationHelper` getätigt werden. Bestehende Anfragen können innerhalb dieser Klassen geändert werden, ohne in die komplexeren Funktionen der Aufarbeitung der Datensätze eingreifen zu müssen. `MASEntityHelper` enthält Hilfsmethoden für die Entity-Tabellen der MAS-Datenbank und `MASRelationHelper` enthält Hilfsmethoden für die Relation-Tabellen. Das Datenbankschema wurde bereits im Abschnitt 4.2 erläutert.

Aufarbeitung der MAS-Datensätze

Zur Aufarbeitung der Datensätze wird wie erwähnt eine HyperSQL In-Memory-Datenbank verwendet. Die zugehörigen Entitäten befinden sich im Paket `de.kbs.acavis.integration.mas.model`. Bis dato werden ausschließlich Daten innerhalb der Klasse `CombinedPaper` dieses Pakets an die Service-Schicht weitergereicht, deshalb wird die dafür notwendige abstrakte Oberklasse direkt implementiert. Die Entities und Relations aus der MAS-Datenbank werden einzeln in die In-Memory-Datenbank eingepflegt, aber nur durch die Klasse `CombinedPaper` ausgelesen.

Das Ein- und Auslesen der Datensätze in und aus der In-Memory-Datenbank werden in der Klasse `InMemoryDao` des Pakets `de.kbs.acavis.integration.mas.controller` vollführt. In eben dieser Klasse wird auch die zugehörige Datenbankverbindung eröffnet. Die Konfiguration der Datenbank befindet sich in der Datei `/src/main/resources/META-INF/persistence.xml`. Zuletzt ist anzumerken, dass kurz nach Eröffnen der Verbindung zur In-Memory-Datenbank sämtliche Fremdschlüssel-Constraints deaktiviert werden. Dies ist nötig, um die im Abschnitt 4.2 besprochenen Probleme mit „illegalen“ Fremdschlüsseln zu umgehen.

Hinzufügen eines weiteren Informationssystems

AcaVis ist vorbereitet, mehrere externe Informationssysteme parallel und selektiv zu betreiben. Eine solche Implementierung eines Informationssystems wird in diesem Absatz als *Integration* bezeichnet, nicht zu verwechseln mit der Integration-Schicht. Die entsprechenden Vorkehrungen in den Schichten Service und Presentation wurden bereits getroffen. Eine Integration kann hinzugefügt werden, in dem zunächst ein entsprechendes neues Paket im Paketpfad `de.kbs.acavis.integration` angelegt wird, siehe als Beispiel die MAS-Integration. Die Schnittstellen zur Service-Schicht werden durch je eine Klasse pro Integration im Paket `de.kbs.acavis.integration.controller` implementiert. Es ist also eine solche Klasse für die neue Integration zu erstellen. Um von AcaVis erkannt zu werden, muss diese die Annotation `@IntegratedInfrastructure` besitzen und die gewünschten Schnittstellen implementieren. Die derzeit möglichen Schnittstellen sind

im Paket `de.kbs.acavis.integration` zu finden und besitzen das `Suffix Provider` im Namen.

Ist eine Integration nicht in der Lage alle einzelnen Methoden einer Schnittstelle zu erfüllen, so kann dies mittels einer `UnimplementedFeatureException` signalisiert werden. Die temporäre oder dauerhafte Nicht-Verfügbarkeit einer Integration wird durch eine `IntegrationUnavailableException` signalisiert. Generell sollten die Implementierungsdetails einer Integration nur im zu Beginn erstellten Paket enthalten sein. Der zusätzlich erstellte Controller, welcher auch die Schnittstellen implementiert, soll lediglich verfügbare Schnittstellen an höhere Schichten kommunizieren und die Informationen „durchleiten“. Die Implementierung der neu erstellten Integration kann in beliebiger Art und Weise mit externen Bibliotheken erfolgen.

Sammlungen, Visualisierungen und archivierte Suchen

Die weiteren Funktionen von AcaVis sind in der Service-Schicht enthalten. Wie erwähnt werden diverse Informationen in einer integrierten Datenbank gespeichert. Die zugehörigen Entitäten der Datenbank sind im Paket `de.kbs.acavis.service.model.persistence` zu finden. Die bereitgestellten Methoden dieser Schicht implementieren die in der Presentation-Schicht benötigten Daten je nach Bedarf. Es muss darauf geachtet werden, diese synchron zu halten. Wichtige Controller-Klassen dieser Schicht sind im Paket `de.kbs.acavis.service.controller` zu finden: `SearchManager` enthält alle notwendigen Funktionen zur Bereitstellung der Suchfunktion, `CollectionManager` stellt alle nötigen Informationen und Methoden zur Verwaltung der Sammlungen bereit und `PublicationManager` beinhaltet Funktionen zur Bereitstellung der Detailansicht.

6.5. Quelltexte und Ressourcen von AcaVis

Die im Rahmen dieser Arbeit entstandenen Quelltexte und Ressourcen von AcaVis werden im Nachhinein kostenfrei unter einer entsprechenden OpenSource-Lizenz zur Verfügung gestellt. Dies soll eine Nutzbarkeit und Weiterentwicklung durch Dritte erleichtern. Die Quelltexte und Ressourcen werden zunächst den Gutachtern dieser Arbeit zur Verfügung gestellt. Nach einer weitergehenden Bereinigung des Projekts sowie einer Ergänzung von Kommentaren und Dokumentation wird das Projekt beim kostenfreien Projekthosting-Dienst *GitHub*⁷⁸ eingepflegt.

Das Projekt wird unter folgender Adresse erreichbar sein:
<https://github.com/AcaVis>

⁷⁸GitHub <http://github.com/>

7. Fazit und Ausblick

Die vorliegende Arbeit befasste sich mit den Möglichkeiten von Visualisierungen als Hilfsmittel der Recherche von wissenschaftlichen Publikationen. Ziel war eine begründete Auswahl von Visualisierungen für diesen Zweck. Eine umfangreiche theoretische Aufarbeitung von Visualisierungen im Allgemeinen war notwendig, um deren generellen Nutzen aufzuzeigen.

Es konnte auch gezeigt werden, welche herausragende Rolle Visualisierungen bei der Aufarbeitung von umfangreichen Datenbeständen spielen können. Dabei wurden die vielen Falltüren beim Erstellen von Visualisierungen analysiert und mögliche Lösungsansätze aufgezeigt. Um die Notwendigkeit eines verbesserten Ansatzes der Recherche zu untermauern, wurden verschiedene dahingehende Konzepte evaluiert. Im Ergebnis wurden hier zwei verschiedene Konzepte identifiziert: Programme die nur die Visualisierung bibliografischer Daten ermöglichen gegenüber Programmen, die lediglich die Recherche von Publikationen unterstützen.

Der daraufhin in dieser Arbeit entwickelte Ansatz einer Recherche mit unterstützenden Visualisierungen soll eine Verbindung zwischen den beiden Konzepten herstellen. Dabei galt es sicherzustellen, dass erstens kein spezielles Fachwissen zur Benutzung der Visualisierungen notwendig ist, zweitens eine kostenlos verfügbare aber dennoch umfangreiche Datenbasis zur Verfügung steht und drittens grafisch exzellente und aussagekräftige Visualisierungen geschaffen werden. Dieser Ansatz bedient sich insbesondere auch gelungener bestehender Konzepte.

Die Visionen und Darstellungen von Autoren anderer Arbeiten inspirierten in großen Teilen die entwickelten Visualisierungen. Im Wesentlichen führten aber die von mir entwickelten Anwendungsfälle zu der vorliegenden Implementierung. Diese schafft die Möglichkeit, nach Publikationen zu suchen, diese in Sammlungen zu organisieren und zu visualisieren. Das sind die absoluten Kernanforderungen, die im Rahmen der Arbeit identifiziert wurden. Entwickelt als Webanwendung steht AcaVis auf einer Vielzahl von Endgeräten zur Verfügung. Nebenbei wird dadurch die Wartung, Fortentwicklung und Verbreitung erleichtert. Die Benutzeroberfläche von AcaVis ist auf das Wesentliche reduziert, stellt aber dennoch die beschriebenen Funktionen mit reichhaltigen Möglichkeiten zur Verfügung. Die darin enthaltenen Visualisierungen bibliografischer Daten wurden speziell für ihren Einsatzzweck entwickelt und optimiert. Die im theoretischen Teil der Arbeit vorgestellten Konzepte wurden dabei angewendet. Die Visualisierungen sind interaktiv und lassen sich ohne tiefe Kenntnis über die zu Grunde liegenden Konzepte interpretieren. Durch die vollautomatische Darstellung obliegt einzig und allein die Interpretation der Visualisierungen dem Betrachter. Die Aussagekraft der Visualisierungen ist damit generell unabhängig von den speziellen Eigenschaften der visualisierten Publikationen.

Es musste im Verlauf des Projektes aber auch festgestellt werden, dass nicht alle aufgetretenen Probleme zur vollen Zufriedenheit gelöst werden konnten. Dazu gehört vor allem der Zugriff auf eine umfangreiche kostenlose Datenbasis. Die aufgezeigten finanziellen Interessen und lizenzrechtlichen Probleme an und mit diesen Daten stellen eine große Herausforderung dar. Auch der Fortbestand der hier verwendeten Datenbasis von Microsoft Academic Search ist nicht gesichert. Es bleibt derzeit keine andere Möglichkeit, als diese Gegebenheiten hinzunehmen und bestmöglich zu berücksichtigen. Der hier entwickelte Ansatz ist ein erster Schritt zur Unterstützung der Recherche mittels Visualisierungen. Der dazu entwickelte Prototyp AcaVis dient der Verbesserung der allgemeinen Situation und soll darüber hinaus zur Fortentwicklung des Ansatzes motivieren.

Weiterführenden Arbeiten bietet sich eine Vielzahl von Möglichkeiten an das Projekt anzuknüpfen. Völlig unabhängig voneinander können Datensammlung, Visualisierungen und Funktionen von AcaVis entwickelt werden. So lassen sich vielerlei Ansätze finden. Einige Beispiele seien hier genannt.

Von besonderer Wichtigkeit im Hinblick auf die Probleme mit der Datenbasis ist es, Alternativen zu Microsoft Academic Search noch eingehender zu untersuchen und einzubinden. Eine Aufgabe mit viel Potenzial ist hier auch die Verknüpfung der Informationen von verschiedenen Informationssystemen, um das Informationsangebot in AcaVis noch reichhaltiger zu gestalten.

Sinnvoll wäre es zudem, weitere Visualisierungen in AcaVis einzubringen oder bestehende Visualisierungen zu verfeinern. Die theoretische Basis dafür können etwa die Darstellungen in dieser Arbeit oder ganz eigene Überlegungen liefern. Ein besonders interessanter Ansatz wäre die Substituierung der traditionellen Zitationsanalyse. Eine Möglichkeit dafür, die auch in dieser Arbeit skizziert wurde, bietet die Analyse von Statistiken über die Lesergemeinde von Publikationen. Solche Statistiken werden zum Beispiel von Mendeley bereitgestellt.

Funktionale Erweiterungen von AcaVis könnten beispielsweise eine erweiterte Suche nach Konferenzen, Journalen oder Institutionen umfassen. Eine Erweiterung der Möglichkeiten der individuellen Arbeit mit Sammlungen und Publikationen ist denkbar. Genannt seien zum Beispiel Notizen, Verschlagwortung und allgemeine kollaborative Ansätze.

Es sei auch noch die Möglichkeit angemerkt, Teilaspekte aus AcaVis in ein anderes bestehendes System zu übertragen. Mittelfristiges Ziel sollte aber sein, AcaVis selbst vom Prototypen zur produktiv einsetzbaren Webanwendung weiterzuentwickeln und so möglichst vielen (angehenden) Wissenschaftlern wertvolle Unterstützung in ihrer Recherchearbeit zu bieten. Mit dieser Arbeit ist der Grundstein dazu gelegt.

Literaturverzeichnis

- Allendoerfer, K., Aluker, S., Panjwani, G., Proctor, J., Sturtz, D., Vukovic, M., und Chen, C. (2005). Adapting the cognitive walkthrough method to assess the usability of a knowledge domain visualization. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 195–202.
- Beel, J. und Gipp, B. (2009). Google scholar’s ranking algorithm: The impact of citation counts (an empirical study). In *Research Challenges in Information Science, 2009. RCIS 2009. Third International Conference on*, pages 439–446.
- Beel, J. und Gipp, B. (2010). Academic search engine spam and google scholar’s resilience against it. *Journal of Electronic Publishing*.
- Bertin, J. (1983). *Semiology of Graphics*. University of Wisconsin Press, Madison, WI.
- Chatfield, C. (1986). Exploratory data analysis. *European Journal of Operational Research*, 23(1):5 – 13.
- Chen, C. (2006). Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377.
- DESTATIS – ohne Autor (2014a). Aus- und Einfuhr (Außenhandel): Deutschland, Jahre, Warensystematik.
- DESTATIS – ohne Autor (2014b). Bierabsatz 2013 weiter rückläufig.
- Elmqvist, N. und Tsigas, P. (2007). Citewiz: a tool for the visualization of scientific citation networks. *Information Visualization*, 6(3):215–232.
- Fry, B. (2008). *Visualizing data - exploring and explaining data with the processing environment*. O’Reilly.
- Hull, D., Pettifer, S. R., und Kell, D. B. (2008). Defrosting the digital library: Bibliographic tools for the next generation web. *PLoS Comput Biol*, 4(10):e1000204.
- Humboldt Institut für Internet und Gesellschaft – ohne Autor (2014). ALADIN: Automated Literature Analysis and Discourse Identifier Network. A set of slides received from Dr. Sascha Friesike.
- Khalid, A., Afzal, M., und Qadir, M. (2011). Citation network visualization of citeseer dataset. In *Computer Sciences and Convergence Information Technology (ICCIT), 2011 6th International Conference on*, pages 367–370.

- Kraker, P., Jack, K., Schloegl, C., Trattner, C., und Lindstaedt, S. (2013). Head start : Improving academic literature search with overview visualizations based on readership statistics. New York, NY. ACM.
- Lohmann, S., Ziegler, J., und Tetzlaff, L. (2009). Comparison of tag cloud layouts: Task-related performance and visual exploration. In Gross, T., Gulliksen, J., Kotzé, P., Oestreicher, L., Palanque, P. A., Prates, R. O., und Winckler, M., editors, *Proceedings of INTERACT (1)*, volume 5726 of *Lecture Notes in Computer Science*, pages 392–404. Springer.
- Modjeska, D., Tzerpos, V., Faloutsos, P., und Faloutsos, M. (1996). Bivteci: A bibliographic visualization tool. In *Proceedings of the 1996 conference of the Centre for Advanced Studies on Collaborative research*, page 28. IBM Press.
- Orduña-Malea, E., Ayllon, J. M., Martín-Martín, A., und López-Cózar, E. D. (2014). Empirical evidences in citation-based search engines: Is microsoft academic search dead?
- Rosvall, M. und Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal al community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123.
- Rosvall, M. und Bergstrom, C. T. (2010). Mapping change in large networks. *PLoS ONE*, 5(1):e8694.
- Sci2 Team - ohne Autor (2009). Science of Science (Sci2) Tool.
- Tufte, E. R. (1986). *The visual display of quantitative information*. Graphics Press, Cheshire, CT, USA.
- van Eck, N. J. und Waltman, L. (2007). Vos: A new method for visualizing similarities between objects. In Decker, R. und Lenz, H.-J., editors, *Advances in Data Analysis, Studies in Classification, Data Analysis, and Knowledge Organization*, pages 299–306. Springer Berlin Heidelberg.
- van Eck, N. J. und Waltman, L. (2010). Software survey: Vosviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2):523–538.
- van Eck, N. J. und Waltman, L. (2014). Citnetexplorer: A new software tool for analyzing and visualizing citation networks. *CoRR*, abs/1404.5322.
- Waltman, L., van Eck, N. J., und Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *CoRR*, abs/1006.1032.

Danksagung

Zunächst danke ich meinem Betreuer Professor Robert Jäschke, für die inspirierenden Gespräche und die angenehme Betreuung.

Dann möchte ich mich bei meinen Eltern bedanken, für die Unterstützung während meines bisherigen Studiums, während meines gesamten Werdegangs und insbesondere während dieser Bachelorarbeit.

Weiterhin gilt mein Dank speziell den folgenden Personen:

Meinem Vater, für die inhaltliche und stilistische Korrektur der Ausarbeitung

Martin Singer, für die kreativen Fachgespräche und inhaltliche Anregungen

Philipp Horst, für die Korrektur des englischsprachigen Abstracts der Ausarbeitung

Anhang

A. Weitere Screenshots der Benutzeroberflächen

Um einen umfangreichen Überblick über die praktischen Ergebnisse dieser Arbeit auch in schriftlicher Form zu geben, sind im Folgenden die Benutzeroberflächen von AcaVis als Screenshots eingebunden. Die wichtigsten Benutzeroberflächen wurden Bereits im Kapitel 5 als Screenshots eingebunden. Ebenso sind dort die Screenshots der Prototypen von Visualisierungen in AcaVis zu finden. Die hier im Anhang enthaltenen Screenshots bilden diverse erweiterte Funktionalitäten ab.

A.1. Suchfunktion

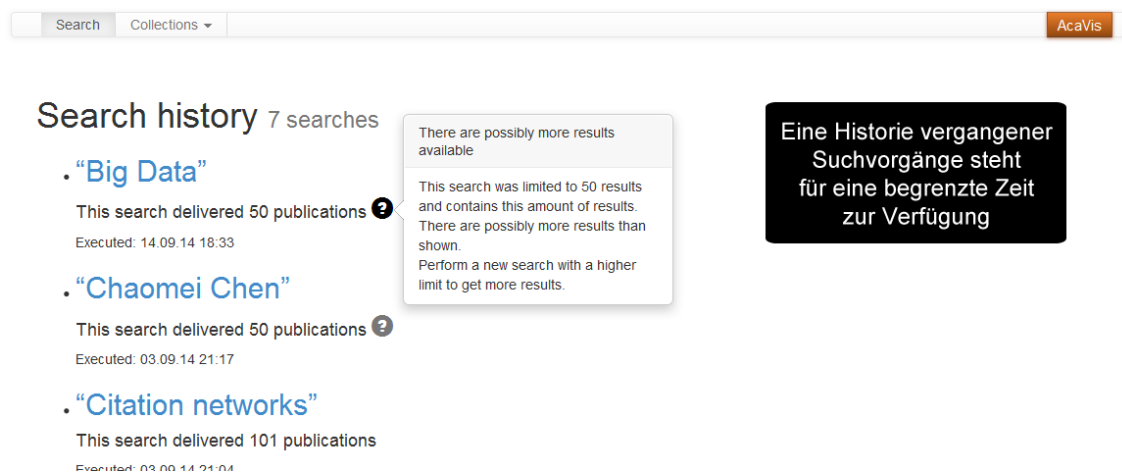


Abbildung A.1.: Getätigte Suchvorgänge werden für einen gewissen Zeitraum archiviert und können aus dem Cache von AcaVis abgerufen werden.

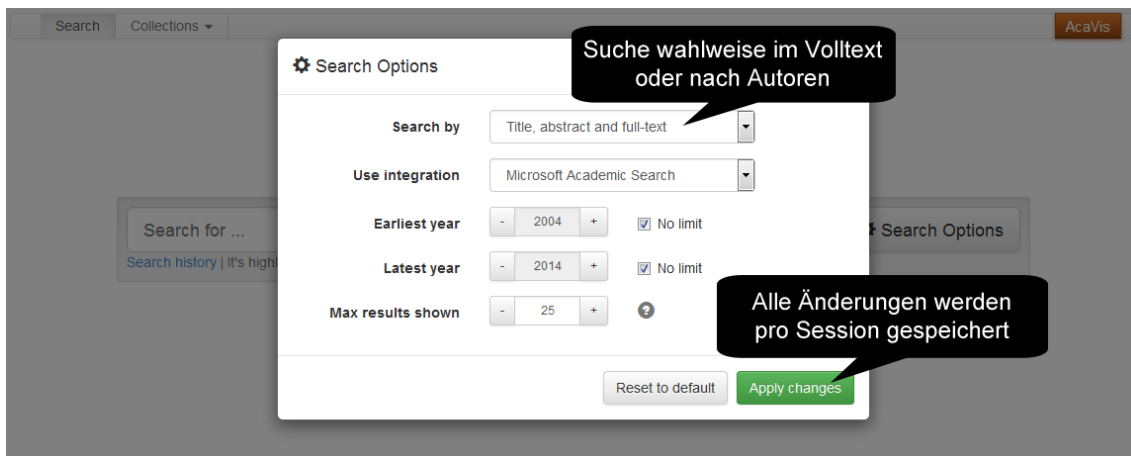


Abbildung A.2.: Im Rahmen einer Suche stehen vielfältige Einstellungsmöglichkeiten zur Verfügung. Viele wurden bereits im Hauptteil der Ausarbeitung erläutert. Die Einstellungen werden in der Session eines jeden Benutzers abgespeichert.

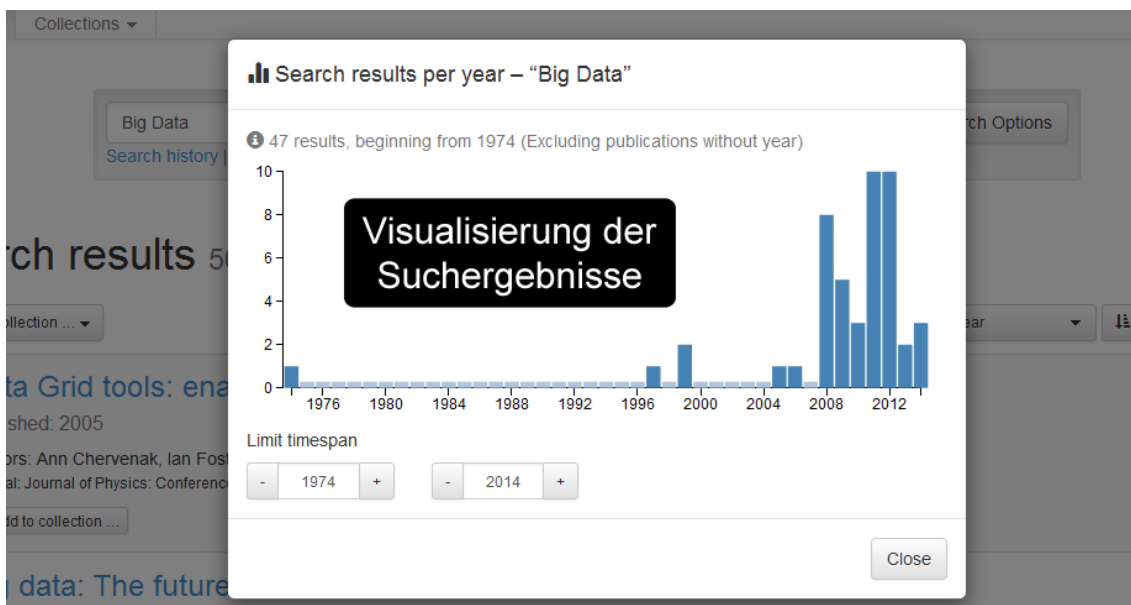


Abbildung A.3.: Eine histogrammische Darstellung von Suchergebnissen.

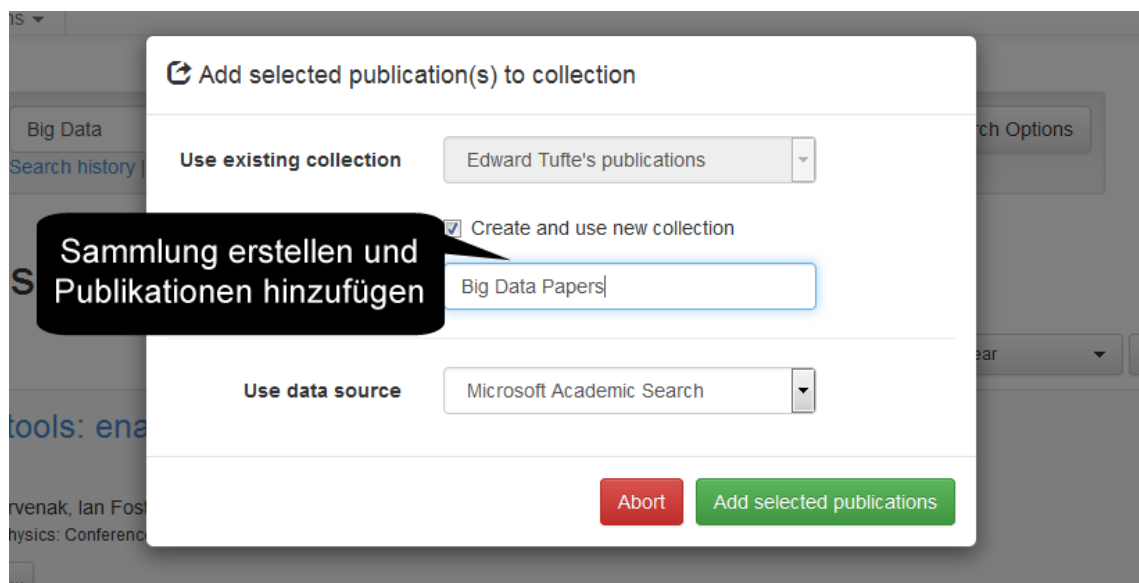


Abbildung A.4.: Publikationen können innerhalb der Suchergebnisse ausgewählt und einer Sammlung hinzugefügt werden. Neue Sammlungen können „On-the-fly“ erstellt werden. Bestehende Sammlungen können ausgewählt werden.

A.2. Sammlungen von Publikationen

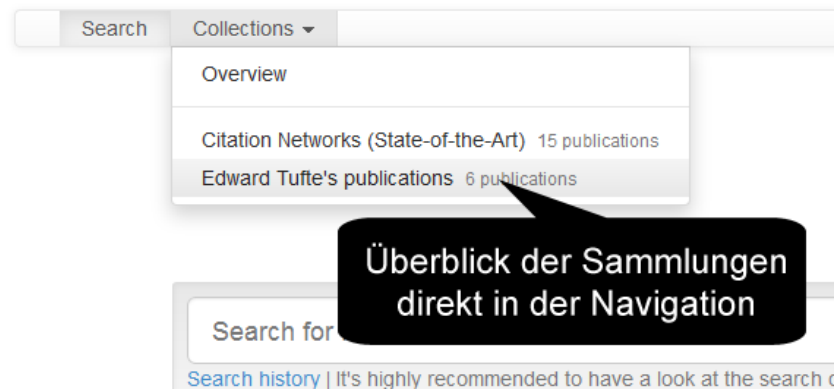


Abbildung A.5.: In der Navigation sind alle verfügbaren Sammlungen zu finden. Die entsprechende Übersicht, wie im Abschnitt 5.4 vorgestellt, kann durch die Navigation aufgerufen werden.

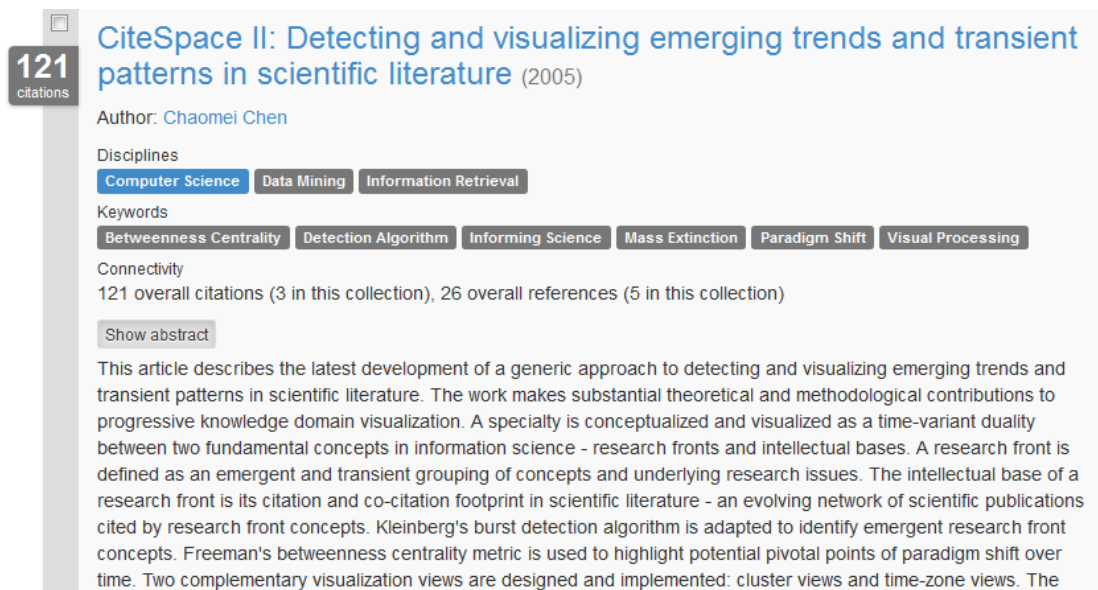


Abbildung A.6.: Um die Übersichtlichkeit zu wahren, werden in der Sammlung zunächst nur die wichtigsten Informationen gezeigt. Weitere Informationen, zum Beispiel das Abstract (falls verfügbar), können bei Bedarf angezeigt werden. Auch die Navigation zur Übersichtsseite einer Publikation ist durch den entsprechenden Link (Publikationstitel) möglich.