Netzwerkanalysen und semantische Datenmodellierung als heuristische Instrumente für die historische Forschung

Der Technischen Fakultät
der Friedrich-Alexander-Universität

Erlangen-Nürnberg zur
Erlangung des Doktorgrades Dr.-Ing.
vorgelegt von
Dirk Wintergrün
aus Düsseldorf

Als Dissertation genehmigt
von der Technischen Fakultät
der Friedrich-Alexander-Universität Erlangen-Nürnberg
Tag der mündlichen Prüfung: 18.2.2019

Vorsitzende/r des Promotionsorgans: Prof. Dr. Reinhard Lerch

Gutachter: Prof. Dr. Günther Görz
Prof. Dr. Klaus Meyer-Wegener
Prof. Dr. Manfred Laubichler

Summary

Today we are facing a transition from the age of information to the age of the networks. The consequence is a revolutionary change in the organisation of knowledge and the structures of its preservation. This process is simultaneously a challenge to researchers working in computer science and in the humanities. Promising answers to these challenges can increasingly be found in two approaches: the semantic modelling of data and the mathematical analysis of networks. The problem is that until now these approaches have been mostly explored separately, although both aim to formalise knowledge structure and systems of knowledge diffusion so that these can be analysed with methods developed by computer scientists.

This thesis brings these two approaches together and outlines a theory for a combined network analysis and model-based description of historical knowledge structures (NMD). Starting from the approach of historical epistemology, this theory introduces three interconnected layers of networks: the semantic network, describing the structures of knowledge; the semiotic network, representing the physical and formal representation of knowledge; and finally the social network of actors, which are indispensable for the structuring and restructuring of knowledge.

Part One of this thesis introduces the theoretical underpinning of the thesis, based on mathematics, graph-theory and modelling, and grounded in the history of science, which is necessary in order to be able to apply these approaches to research in this area. Part Two is devoted to the concrete realisation of this approach, and introduces four case studies. Drawing on examples from four different historical periods, and addressing four different research questions, it is demonstrated which tools can be used to answer these questions and which consequences follow from NMD for interdisciplinary historical research. The case studies are projects researching a) the history of general relativity (GR), b) the history of the Max Plank Society (GMPG), c) the construction of the Cuppola of the Florentine Cathedral, and d) the Sphaera of Sacrobosco.

The process of designing a workflow for data transformation and data modelling is at the center of both the project on the history of general relativity (GR) and the study based on data from the project on the history of the Max Planck Society (GMPG). This workflow allows to bring together different data sets based on semantic modelling, so that this data can be analysed using the tools of social network analysis. The conceptual relations between semiotic, semantic and social network are then discussed, based on the example of bibliometric studies and social networks in the case study of GR. The case study on GR also focuses in particular on the effects of missing or uncertain data in the results of network analysis. The case study GMPG highlights how institutions (in this case commissions) can be interpreted as elements of a semiotic networks in this theoretical framework. The studies on the cuppola of the Florentine cathedral and the sphere of Sacrobosco focus on the connection of semiotic networks (represented by archival materials and books), the semantic network, and partially the social network. An ontology which supports this approach is presented here.

In conclusion, the approach presented here is highly useful for making computational methods more accessible to humanists. Quantitative data also becomes more understandable for researchers who are predominantly interested in the qualitative interpretation of data.

Zusammenfassung

Stand das 20. Jahrhundert mit dem Aufkommen der elektronischen Medien und dem Internet als Massenphänomen für eine grundlegende Umwälzung der Informationsvermittlung und -organisation, so befinden wir uns nun in einer Umbruchphase vom Informations- zum Netzwerkzeitalter. Diese konfrontiert uns mit einer grundlegenden Veränderung der Organisation des Wissens und den Strukturen seiner Speicherung. Dieser Umbruch ist eine Herausforderung an die Informatik und an die Geisteswissenschaften gleichermaßen. Vielversprechende Antworten auf diese Herausforderungen ergeben sich aus meiner Sicht vor allem aus zwei Ansätzen - der semantischen Modellierung und der mathematischen Analyse von Netzwerken. Diese methodischen Ansätze werden jedoch bisher zumeist getrennt voneinander behandelt, obwohl beide Ansätze dazu eingesetzt werden, Wissen zu strukturieren und zugleich maschinenlesbar zu machen. Wissensstrukturen werden damit einer Auswertung mit Mitteln der Informatik zugänglich. Die Verbindung dieser beiden Ansätze ist der Leitgedanke dieser Arbeit. Notwendig ist eine Theorie zur netzwerk- und modellierungstheoretischen Beschreibung (NMB) historischer Wissenssysteme. Diese Theorie basiert ausgehend vom wissenschaftshistorischen Ansatz der historischen Epistemologie auf drei miteinander verschränkten Netzwerkebenen: dem semantische Netz, das die Wissensbasis beschreibt, dem semiotischen Netz, das die Kodierung von Wissen repräsentiert, und schließlich dem Netz der sozialen Akteure, ohne die Entstehung und Organisation von Wissen nicht stattfinden kann. In Teil I der Arbeit wird in die mathematischen, graphentheoretischen und modellierungstheoretischen Ansätze soweit eingeführt, dass die Umsetzung in konkrete Anwendungen verständlich ist, gleiches gilt für den notwendigen Hintergrund von Seiten der Wissenschaftsgeschichte.

Teil II der Arbeit ist der konkreten Umsetzung in vier Fallstudien gewidmet. Anhand von Beispielen aus unterschiedlichen historischen Zeitabschnitten verbunden mit unterschiedlichen Fragestellungen wird aufgezeigt, welche Hilfsmittel eingesetzt werden können, um diese Fragestellungen zu beantworten, und welche Konsequenzen sich aus dem Ansatz der NMB für die interdisziplinäre historische Forschung ergeben. Im Zentrum stehen bei den Fallstudien zur Geschichte der Allgemeinen Relativitätstheorie (ART) und dem Teilprojekt aus dem Projekt zur Geschichte der Max-Planck-Gesellschaft (GMPG) der Aufbau einer Infrastruktur, die auf Grundlage einer semantischen Modellierung unterschiedliche Datenbestände zusammenbringt und diese dann einer soziale Netzwerkanalyse zugänglich macht. Als Beispiel für die Verbindung von semiotischem, semantischem und sozialen Netz wird in der Fallstudie zur ART die Verbindung zwischen bibliometrischer Analyse und dem Kooperationsnetzwerk der beteiligten Wissenschaftlerinnen und Wissenschaftler vorgestellt. Wir untersuchen hier zugleich die Auswirkungen von historischen Annahmen sowie ungenauer oder fehlender Daten auf die Ergebnisse der Netzwerkanalyse. In der Fallstudie zur GMPG schauen wir darüber hinaus darauf wie Gremien (in diesem Falle Kommissionen), als semiotische Netzwerke verstanden werden können. Die Studien zum Bau der Kuppel des Florentiner Doms und der Sphaera des Sacrobosco legen den Schwerpunkt auf die Verbindung von semiotischem Netz repräsentiert durch Schriftzeugnisse in Archiven und semantischem Netz. Es wird eine beispielhafte Ontologie vorgestellt, die dieses unterstützt.

Die Ergebnis dieser Arbeit verdeutlichen, dass der dargestellte Ansatz dabei hilft, Methoden der Informatik für die geisteswissenschaftliche Forschung zugänglicher zu machen, quantitative Datenauswertung wird dadurch besser verständlich auch für vorrangig an qualitativen Ergebnissen interessierten Forscherinnen und Forschern aus den Geisteswissenschaften. Es ist mit vertretbarem Aufwand möglich historische Datenbestände mit Methoden der NMB aufzubereiten und Fragen an diese Daten, wie etwa nach Abhängigkeitsbeziehungen oder der Relevanz einzelner Personen in historischen Prozessen zu stellen und zu beantworten. Die Arbeit zeigt, dass dazu keine monolithische neue Infrastruktur notwendig ist, sondern dass durch flexible Kombination existierender Methoden und deren Übertragung auf neue Anwendungsbereiche bereits erhebliche Fortschritte und neue Erkenntnisse erzielt werden können.

Inhaltsverzeichnis

Kapitel 1

Einführung

Warum diese Arbeit und warum zu diesem Zeitpunkt? Wir befinden uns im Zentrum einer globalen Umwälzung der Wissensorganisation. War das 20. Jahrhundert verbunden mit dem Aufkommen der elektronischen Medien und dem Internet als Massenphänomen mit einer grundlegenden Umwälzung der Informationsvermittlung und -organisation, so sehen wir uns nun mit einer grundlegenden Veränderung der Organisation und Strukturierung von Wissen konfrontiert. Verstehen wir Daten als die fundamentalen Bausteine des Internet, Information als die Verknüpfung dieser Daten und schließlich Wissen als inkorporierte Handlungsoptionen und Strukturierung dieser Informationen [262], so ist es offensichtlich, dass sich die Geisteswissenschaften, in deren Zentrum die Struktur und Geschichte menschlichen Wissens und Handelns stehen, mit Netzwerken und der Modellierung von Informationen in Netzwerken befassen müssen.[1] Dieses gilt umso mehr, als das Netzwerk der Daten nicht losgelöst von gesellschaftlichen Verflechtungen und den immer mehr miteinander kommunizierenden Alltagsgegenständen gedacht werden kann. Die Wissensbasis des Einzelnen in der vernetzten Gesellschaft konstituiert sich über seine Einbindung in das Netz. Wir befinden uns in einer Umbruchphase vom Informations- zum Netzwerkzeitalter. Es muss den Geisteswissenschaften darum gehen, die in komplexen Abhängigkeitsverhältnissen stehenden Teilnetze, die dieses neue Zeitalter prägen, zu verstehen und zu beschreiben. Dies ist ohne ein zumindest prinzipielles Verständnis der hinter den Netzwerken und ihrer Beschreibung stehenden Technologien und Methoden nicht zu leisten und eine der wichtigsten Herausforderungen an die interdisziplinäre Kooperation von Geisteswissenschaften und der Informatik.

Geisteswissenschaftlerinnen und -wissenschaftler müssen in die Lage versetzt werden, die technischen Methoden und dahinter stehenden Algorithmen verstehen und nachvollziehen zu können. Zugleich wird unter dem Blickwinkel des Netzwerkzeitalters das Netz zu mehr als einer technischen Herausforderung, es wird unmittelbar zu einer Frage der Wissensgenerierung und -strukturierung und damit zu einem Problem der Epistemologie. Die Informatik hat sich hier ihrer Funktion zu stellen, eine der tragenden Säulen für die Codierung und Überlieferung menschlichen Wissens geworden zu sein. Dafür bedarf es eines gleichberechtigten Dialoges zwischen Informatik und Geisteswissenschaften.

[1] In der Auseinandersetzung mit dem Informationsbegriff und seiner Bedeutung für die Wissensgeschichte hat sich diese Definition von Information aus unser Sicht als fruchtbar für die weitere Einordnung in netzwerktheoretische Betrachtungen ergeben. In [204] wird dieses als Ergebnis eines gemeinsamen Diskussionsprozesses weiter ausgeführt.

Es ist notwendig, die Grundlage für eine gemeinsame Sprache dieser Disziplinen zu legen. Diese Arbeit will einen Beitrag dazu leisten. Hierbei konzentriere ich mich auf zwei Ansätze, die zunehmend als ein vielversprechender Weg hin zu einer computergestützten Geisteswissenschaft gesehen werden: Die semantische Modellierung und die graphentheoretische Beschreibung von Netzwerken. Diese methodischen Ansätze werden jedoch bisher zumeist getrennt voneinander behandelt, obwohl beide Ansätze darauf abzielen, Wissensstrukturen, die im Zentrum der geisteswissenschaftlichen Forschung stehen, systematisch in maschinenlesbarer Form abzubilden und damit eine Auswertung mit Mitteln der Informatik zu eröffnen. Zentraler Aspekt dieser Arbeit ist hierbei, dass eine Verbindung dieser beiden Methoden zu grundsätzlich neuen Herangehensweisen in den Geisteswissenschaften führt und zugleich neue Forschungsfelder der Informatik erschließt. Die semantische Modellierung erlaubt eine präzisere Definition der Knoten und Kanten eines Netzwerkes und erleichtert damit eine Interpretation der Ergebnisse der mathematischen Graphen- bzw. Netzwerktheorie. Den Geisteswissenschaften ist dieser Ansatz mit nur geringen Übersetzungskosten vermittelbar, da er nahe an der Struktur der dort vertrauten Argumentationsmuster und Organisationsprinzipien von wissenschaftlicher Erkenntnis liegt.

Aus geisteswissenschaftlicher Sicht stellt sich die Herausforderung, das Wissenssystem „Netz“ zu verstehen und die sich daraus ergebenen Konsequenzen für die Ausbildung, für zukünftige Formen der Arbeit, vor allem aber für die demokratische Kontrolle dieses Wissenssystems zu formulieren. Dazu müssen grundlegende Strukturen der sozialen Netzwerktheorie genauer beleuchtet werden. Die Rolle von sozialen Bindungen in einem Netz für die Organisation von Wissensstrukturen – insbesondere für die Expansion und Rekonfiguration dieser Strukturen – ist bisher nur partiell untersucht werden. Das Wechselspiel zwischen Repräsentation von Wissen in Form eines *semiotischen Netzwerkes*, getragen von sozialen Akteuren, mit einem *semantischen Netzwerk* ist eine der Kernfragen für das Verständnis des Wissenssystems „Netz“.

Dies bedeutet, dass eine Soziologie und ökonomische Theorie des Netzes nicht ausreichen kann, um die tiefgreifenden Umwälzungen des Netzwerkzeitalters zu erfassen. Eine intensive Erforschung des Wechselspiels von sozialen Netzwerken und dem mit ihnen in Korrespondenz befindlichen Wissenssystem ist unumgänglich: Es bedarf einer Epistemologie des Netzes.

Dabei kann es jedoch nicht nur um eine theoretische Reflexion gehen, sondern es muss auch der Nachweis angetreten werden, dass diese Methoden auch in der praktischen Arbeit in historischen Forschungsprojekten eingesetzt werden können und dort zu neuen Einsichten führen. Daher sind, neben der theoretischen Reflexion, die Darstellung praktischer Arbeitsabläufe in der historischen Forschung sowie die Vorstellung und Bewertung konkreter Entwicklungen von Arbeitsumgebungen Teil dieser Arbeit. Der Schwerpunkt liegt auf den Komponenten zur Netzwerkanalyse und Datenmodellierung. Dies wird im praktischen Teil der Arbeit vorgestellt, in dem sowohl die Zusammenarbeit innerhalb eines interdisziplinären Teams mit dem Ziel, computergestützte Methoden zum Einsatz zu bringen, erläutert wird, als auch grundlegende Bibliotheken und Anwendungen vorgestellt werden, die als Open-Source-Pakete nachgenutzt werden können und Teil einer Arbeitsumgebung für die digitalen Geisteswissenschaften am Max-Planck-Institut für Wissenschaftsgeschichte (MPIWG) sind, deren Prototyp dort zurzeit entwickelt wird.

1.1 Graphen und Netzwerke – Herausforderung für die Geisteswissenschaften und die Informatik

Graphen im Sinne der mathematischen Graphentheorie stellen für die Informatik[2] seit ihrem Beginn ein grundlegendes Konzept dar. Graphen treten als Problemstellung von außen in die Informatik hinein: Es sei hier an das Travelling-Salesman-Problem [102] oder die Optimierungen von Netzwerken, wie etwa den Netzen der Energieversorger, erinnert. Zugleich sind Graphen ein Ansatz, an grundlegende theoretische Fragen der Informatik heranzugehen, so etwa bei Problemen in der Künstlichen Intelligenz, wie assoziativen Netzwerken, semantischen Strukturen, probabilistischen Schlussfolgerungsnetzen oder schließlich neuronalen Netzen [93, S.259ff] [93, S.357ff].

Die Fragen, die aus den Geisteswissenschaften an die Informatik herangetragen werden, ergeben sich zunächst häufig aus dem sehr pragmatischen Wunsch, Strukturen näher zu untersuchen, die sich auf Grund ihrer Größe und Komplexität nicht mehr ohne technische Unterstützung analysieren lassen. Immer größere Quellenbestände liegen digital vor, sei es in Form von digitalisierten Quellen oder digital erzeugten Ressourcen, die eine ausschließlich digitale Existenz haben. Die Herausforderung, die sich hier konkret stellt, besteht darin, Algorithmen zu entwickeln, die dabei helfen, die Struktur und die Dynamik von Netzwerken besser zu verstehen. Die Aufgaben sind hierbei die Reduktion der Komplexität großer Netzwerke sowie die Bestimmung von quantitativen Größen, mit deren Hilfe sich Netzwerke einordnen und vergleichen lassen. Die Methoden, die zur Anwendung kommen, umfassen einfache statistische Methoden und reichen bis hin zur Anwendung von Ergebnissen aus der Theorie komplexer Systeme. Daneben wird ein anderer Bereich der Informatik, die Informationsvisualisierung in Form der „Visual Analytics“, als eine neue Form der Publikation von historischen Forschungsergebnissen immer populärer.[3]

In der sozial- und wirtschaftswissenschaftlichen Forschung finden die soziale Netzwerkanalyse und die Erforschung der Verbreitung von Innovationen in Diffusionsnetzwerken eine immer breitere Anwendung [246, 122]. Die historische Forschung nimmt solche Ansätze im Rahmen der historischen Netzwerkforschung auf. Bisher liegt hier der Schwerpunkt im Wesentlichen auf der Anwendung und Übertragung der Methoden der sozialen Netzwerkanalyse auf die historische Forschung.[4]

Bislang wird zumeist nur der im engeren Sinne mathematische Teil der Graphentheorie auf die Anwendbarkeit in der historischen Forschung untersucht und für diese fruchtbar gemacht. Für die historische Forschung ist jedoch auch ein weiterer Aspekt der Anwendung von Graphen von großem Interesse, nämlich die Darstellung von Informationen in Form von Wissensgraphen, die sich mit Hilfe semantischer Modellierung strukturieren und formal in RDF bzw. OWL darstellen lassen. Ein roter Faden in dieser Arbeit besteht darin zu zeigen, dass durch die Verbindung von sozialer Netzwerkanalyse und semantischer Modellierung ein Instrumentarium geschaffen werden kann, das dabei hilft, die historisch-kritische Methode der Geschichtswissenschaften auf netzwerktheoretische Überlegungen zu übertragen. Umgekehrt ergeben sich für die Informatik neue Ansätze für die Strukturierung und

[2] Für eine Einführung in die Anwendungsfälle siehe [101, Kap.20].

[3] Ich werde in Abschnitt 5.12.4 näher darauf eingehen. Eine gute Einführung in den Fragenbereich liefert [129].

[4] Einen guten Einblick liefert das kurze aber sehr informative Handbuch zur historischen Netzwerkforschung [68].

Vorhaltung von Wissensdatenbanken.[5].

1.1.1 Informatik: Gegenstand und Partner der Geisteswissenschaften

Die Digitalisierung bedingt einen Methodenwandel in den Geisteswissenschaften, der in ihren einzelnen Disziplinen mit sehr unterschiedlicher Geschwindigkeit und Tiefe voranschreitet. Für die Geschichtswissenschaften heißt dies, dass die etablierten Methoden der detaillierten akribischen Einzelstudie mit den Methoden, die sich aus der computergestützten Auswertung von Massendaten ergeben, zusammengeführt werden müssen. Es gilt, Methoden zu entwickeln, die eine kritische Bewertung auch der durch computergestützte Methoden gewonnenen Ergebnisse ermöglichen. Neben der notwendigen Offenheit gegenüber diesen Methoden von Seiten der Fachwissenschaftler erfordert dieses auch von Seiten der Informatik die Bereitschaft, sich auf Fragestellungen und Denkweisen der Geisteswissenschaften einzulassen.

Die historischen Wissenschaften, auf die sich diese Arbeit konzentriert, stehen insbesondere vor der Herausforderung, dem Paradigma der Detailstudie ein neues Paradigma einer computergestützten Hermeneutik entgegenzustellen, die den Einzelfall in größere Kontexte setzt. Der Einzelfall wird hierbei nicht entwertet – im Gegenteil kann seine Rolle dadurch gestärkt werden, dass seine Bedeutung im komplexen System historischer Prozesse genauer charakterisiert werden kann. Hierzu ist das Erlernen von Grundtechniken der Informationsverarbeitung auch in der Ausbildung von Historikerinnen und Historikern die eine Voraussetzung, die andere ist eine gemeinsame Sprache, die eine Vermittlung der Ergebnisse zwischen den beiden Welten ermöglicht.

In den Geisteswissenschaften zeichnet sich ein sehr differenziertes Bild ab, wenn wir die Anwendungsbreite digitaler Methoden in der Praxis betrachten. So wenig es einen einheitlichen klassischen Methodenkanon für alle Geisteswissenschaften gibt, umso weniger gilt dies für die Anwendung digitaler Methoden. So ist die Linguistik schon früh auf die Möglichkeiten der Analyse von Sprache und Sprachstrukturen mit Hilfe von Computern aufmerksam geworden. Dies ist sicher auch der Tatsache zuzuschreiben, dass zwischen dem Problem des Parsings von formalen Sprachen und den Strukturen gesprochener und geschriebener realer Sprachen eine doch sehr deutliche Parallele besteht. Umgekehrt besteht aus der Sicht der Informatik ein Interesse am Verständnis der Struktur der lebenden Sprachen, um ausgerichtet an natürlichen Sprachen höhere Programmiersprachen zu entwickeln, die von Anwendern mit sehr unterschiedlichen Sprachhintergründen schnell verstanden und erlernt werden können. Es sei in diesem Zusammenhang auch an die Ansätze des „Literal Programming“ erinnert, in denen die Lesbarkeit und das Verständnis formaler Sprache explizit durch eine Umsetzung in natürliche Sprache verbessert werden soll. [133].

Neben der Strukturanalyse ist das Problem der Indizierung, der Suche und des Sortierens von Werten und Varianten ein Problem, das unmittelbar zu Fragen führt, die mit Hilfe der Informatik angegangen werden können. Mit dem Ziel, Zusammenhänge zwischen Häufigkeiten, Korrelationen und der Relevanz von Entitäten zu verstehen, hat die Statistik in viele Bereiche der Geisteswissenschaften Einzug gehalten.

[5] Ansätze dafür finden sich in in den Arbeiten von Gerd Graßhoff seit Mitte der 1990er Jahre im Kontext seiner Theorie zur Beschreibung wissenschaftlicher Entdeckungsprozesse [99, 54, 98].

In der Archäologie gehört der Computer bereits seit Jahrzehnten zum täglichen Handwerkszeug eines großen Teils der dort wissenschaftlich Arbeitenden [16] – zumeist in Form eines Hilfsmittels zur Erfassung, Verwaltung und Analyse von Artefakten, weniger jedoch als Tool zur Interpretation des Kontextes und zur Entdeckung von Zusammenhängen unterschiedlicher Funde im Sinne einer geistes- bzw. menschheitsgeschichtlichen Interpretation.[6] Die Anwendung digitaler Methoden wird hier nicht zuletzt dadurch erleichtert, dass die grabende Archäologie, wie auch die medizinische Anthropologie, die Restaurationswissenschaften oder auch die Psycholinguistik im Fächerkanon eine hybride Funktion einnehmen, in der naturwissenschaftliche Methoden grundsätzlich zum Handwerkszeug der entsprechenden Disziplin gehören.

Trotz großer bestehender methodischer Probleme sind jedoch — bewusst oder unbewusst — die neuen Medien auch in den historischen Wissenschaften ein fester Bestandteil der täglichen Arbeit. Dominant sind jedoch zumeist die Benutzung von Werkzeugen zur Recherche – und dort überwiegend klassische Suchmaschinen und die online-verfügbaren Nachweissysteme der Bibliotheken, Archive und Museen. In Studien über das Nutzerverhalten unter Historikerinnen und Historikern ist das meist genannte Tool nach wie vor das klassische Textverarbeitungssystem [227, 228]. Computergestützte Methoden zur Auswertung und Analyse von Daten werden nur sehr begrenzt eingesetzt und sind auf noch wenige Projekte beschränkt. Überzeugende Ergebnisse, die mit diesen neuen Methoden erzielt werden, finden nur selten Niederschlag in den klassischen historischen Fachzeitschriften. Auf den Konferenzen der Fachgesellschaften gibt es zwar nun in verstärktem Maße Diskussionspanels und -workshops zu digitalen Methoden, aber nur selten – wenn auch mit steigender Tendenz – finden sich Ergebnisse dieser Methoden in den Publikationen selbst wieder.

Hinter den Kulissen der großen Konferenzen gibt es jedoch eine aktive Gemeinschaft, die sich im Kontext der Digital Humanities, in den digitalen Ablegern der Fachgemeinschaften wie der digitalen Kunstgeschichte, aber auch im Rahmen der Infrastrukturinitiativen wie DARIAH und CLARIN austauscht und Wege sucht, um Methodentransfers zwischen den geisteswissenschaftlichen Disziplinen und zwischen Geistes- und Naturwissenschaften sicherzustellen.

Eine besondere Rolle kann in diesem Vermittlungsprozess die Informatik spielen. Insbesondere die im weitesten Sinne als Semantische Technologien charakterisierbaren Arbeitsfelder bieten optimale Voraussetzungen für eine fruchtbare Zusammenführung von geisteswissenschaftlichen und mathematisch-naturwissenschaftlichen Methoden. So kann zum Beispiel auf eine lange gemeinsame Tradition im Umgang mit Problemen der Logik, insbesondere der Frage nach der Beschreibung und Modellierung von Entscheidungsprozessen, zurückgegriffen werden.

Auf der Seite der historischen Forschung bietet sich umgekehrt die Wissenschaftsgeschichte, insbesondere der als historische Epistemologie zusammengefasste Ansatz [208], der Wissens- und Wissenschaftsgeschichte als Geschichte des Zusammenspiels von Handlungs- und Erfahrungswissen sowie kognitiven Strukturen und deren Kodierung versteht, als Partnerin für die Etablierung eines neuen methodischen Paradigmas, das vielleicht besser mit „Computational Humanities“ als mit „Digital Humanities“ beschrieben werden kann. Auf Grund ihrer Fragestellungen zieht die Wissenschafts-

[6] In diese Richtung gehen Projekte, wie der Atlas der Innovationen [51] oder im Überlappungsbereich von Anthropologie und Archäologie: D-Place [64] oder Pulotu [193].

geschichte grundsätzlich unterschiedliche Quellengattungen – sowohl Texte als auch Artefakte – sehr unterschiedlicher Provenienz heran. Sie verfolgt einen grundsätzlich interdisziplinären Ansatz, der verschiedenste natur- und geisteswissenschaftliche Methoden und Quellen vereinen muss. Zugleich ist die Formierung und Transformation von Wissenssystemen ihr Untersuchungsgegenstand. Die Bedeutung von Wissen und seine Kodierung im Netzwerkzeitalter gerät somit zwangsläufig auch in den Fokus der wissenschaftsgeschichtlichen Untersuchungen. Auch hier zeigt sich eine doppelte Rolle der Informatik: Einerseits sind die Konsequenzen ihrer Anwendung Gegenstand der historischen Forschung, andererseits finden ihre Methoden als ein Hilfsmittel zur Strukturierung und Erschließung der Wissensstrukturen des 20. und 21. Jahrhunderts unmittelbar Eingang in die Forschung. In beiden Fällen ist eine Grundkenntnis der Methoden der Informatik unerlässlich. Die Fallbeispiele werden aufzeigen, wie sich dieses in konkreten Problemen, Anforderungen und Lösungen niederschlägt, und welche Synergien sich aus einer gemeinsamen Arbeit von Informatikern und Historikern ergeben.

Aus der Perspektive des Historikers geht es hier bei um die Etablierung digitaler Verfahren als neue heuristische Instrumente und die Anwendung neuer hermeneutischer Methoden. Aus der Perspektive des Informatikers steht die Entwicklung von Algorithmen und Formalismen, die historischen Denkweisen angemessen sind, im Vordergrund, um als langfristiges Ziel neue Lösungsansätze für Probleme wie etwa bei der Formulierung von unscharfen Aussagen, dem Auffinden von Widersprüchen innerhalb von Aussagesystemen sowie der Modellierung von Handlungsmodellen zu finden.

Methodenwandel in den Geschichtswissenschaften

Auf besonderes Interesse trifft insbesondere in den Geschichtswissenschaften die Anwendung von Methoden zur Strukturierung von Wissen und die formale Analyse dieser Strukturen, die sich aus den Überlegungen zum *semantischen Web* ableiten lassen. Daneben nimmt die aus der sozialwissenschaftlichen Netzwerkforschung abgeleitete historische Netzwerkforschung einen immer breiteren Raum innerhalb des Methodendiskurses ein. Noch haben diese Methoden nicht den Mainstream der historischen Forschung erreicht, doch ihr Klopfen an der Tür des Gebäudes etablierter Forschungsmethoden ist immer lauter zu hören. Dieser Methodenwandel hat auch im Fach selbst bereits eine relativ lange, wenn auch nicht immer von der breiten wissenschaftlichen Öffentlichkeit wahrgenommene Tradition. Nicht nur Pater Busa [125], dessen Arbeiten im Bereich der Sprachwissenschaften mit gutem Recht zum Gründungsmythos der *Digital Humanities* gehören, sondern auch – häufig zu unrecht – weniger bekannte Akteure, wie Peter Damerow, Bob Englund und Hans Nissen [168] in der Wissenschaftsgeschichte bzw. Assyriologie, haben bereits frühzeitig die Stärken von computergestützten Methoden auch in den Geisteswissenschaften erkannt und konkret genutzt, um zu wissenschaftlichen Erkenntnissen zu gelangen, die ohne diese Methoden nicht erreichbar gewesen wären.

Auf der Seite der Quellen stehen dank der Digitalisierungskampagnen der Bibliotheken und Museen, aber auch durch Projekte wie beispielsweise der *Europeana* oder der *Deutschen Digitalen Bibliothek* Bestände bereit, die der geisteswissenschaftlichen Bearbeitung harren. In unzähligen Projekten an Universitäten und Forschungseinrichtungen sind Datenbanken[7] entstanden, die einen noch zu hebenden Schatz für eine digitale Geschichtswissenschaft darstellen, wenn es gelingt, diese zu einer

[7] Zur Definition von Datenbanken in unserem Kontext siehe Abschnitt 4.1.

nachnutzbaren Ressource zusammenzuführen. Der jetzige Stand ist jedoch, dass die verfügbaren Forschungsdaten oftmals spezifisch auf einzelne Forschungsfragen hin zusammengetragen wurden und Anwendungen auf das entsprechende Projekt hin konzipiert wurden. Im besten Fall sind sie im Netz verfügbar und standardisiert abzufragen. Der Normalfall ist jedoch ein Komplex von Daten, die außerhalb des Entstehungs- und direkten Anwendungskontextes nicht oder nur eingeschränkt nachgenutzt werden können. Diese Arbeiten stellen oft Pionierleistungen auf ihren einzelnen Gebieten dar – daher wäre es in hohem Maße ungerecht, dies den beteiligten Wissenschaftlerinnen und Wissenschaftlern vorzuwerfen. Ohne diese Arbeiten wäre die Transformation der Methoden der Geisteswissenschaften in das digitale Zeitalter undenkbar. Der Aspekt, dass Daten, die zum Erreichen einer spezifischen Forschungsfrage gesammelt wurden, in anderen Kontexten nachnutzbar sein könnten und durch Kombination mit anderen Ergebnissen neue Forschungsfragen erschließen, ist jedoch erst sehr spät thematisiert worden, und die Konsequenzen für die Datenhaltung und -modellierung, die sich aus einer solchen Nachnutzung ergeben, sind bis heute nicht ausreichend diskutiert worden. Die in dieser Arbeit vorgestellten Ansätze zur Datenmodellierung sind ein Beitrag in diese Richtung.

Auf europäischer Ebene sind mit den großen Infrastrukturinitiativen DARIAH und CLARIN grundlegende Schritte in Richtung einer einheitlichen digitalen Infrastruktur für die Geisteswissenschaften getan worden. Wohin diese Projekte langfristig führen, und welche Akzeptanz sie in den Geisteswissenschaften finden werden, ist jedoch immer noch unklar und hängt von einer noch zu leistenden Methodenreflexion ab, die eine engere Anbindung der Entwicklungen an die heterogenen Forschungsprozesse der Geisteswissenschaften ermöglicht.[8]

Vorarbeiten liegen nun also zur Genüge vor. Es gibt auch für die historische Forschung einen Reichtum an Daten, dessen Potential nun genutzt werden muss. Von zentraler Bedeutung ist jetzt die Frage, welche grundlegenden Methoden notwendig sind, um die verschiedensten Ansätze zusammenzubringen und für die historische Forschung nicht nur im Rahmen spezieller Einzelstudien nutzbar zu machen, sondern mit der Etablierung einer neuen Kulturtechnik tatsächlich einen Methodenwandel innerhalb der historischen Forschung zu erreichen.

1.1.2 Semantische Modellierung – Herausforderung an die Geisteswissenschaften und die Informatik

Semantische Modellierung ist eine Grundvoraussetzung dafür, das in Datenbanken vorliegende Wissen über eine Zeitperiode oder einen Sachverhalt in nachvollziehbarer Weise darzustellen und computergestützt zu analysieren. Eine aus geisteswissenschaftlicher Anwendersicht verständliche und vertretbare Modellierung kann dazu beitragen, die bisher heterogenen Datenbestände aus unterschiedlichen Projekten zusammenzuführen und zu vereinheitlichen. Damit wird die Voraussetzung für eine Weiterentwicklung des Internet zu einem epistemischen Web gelegt. Daten, die auf der Grundlage einer formalen Modellierung verbunden werden, können mit etablierten Verfahren aus der KI

[8] Aus nachvollziehbaren Gründen stehen beide Projekte in stark disziplinären Kontexten, die eng mit ihrer Entstehungsgeschichte vor allem in Deutschland zusammenhängen. Sie decken die beiden großen Bereiche der Philologie und Linguistik gut ab, wenn auch fokussiert auf die deutsche Sprache, während andere Bereiche der Geisteswissenschaften erst langsam in ihren Fokus rücken.

auf Inkonsistenzen überprüft werden und durch Inferenzmethoden erweitert werden. Auf diese Weise entstehen Netzwerke, die auf unterschiedlichen epistemischen Ebenen Wissen strukturieren, Beziehungen zwischen Akteuren beschreiben und komplexere Handlungsmuster zueinander in Beziehung setzen. Konkret geht es hierbei um Beziehungen zwischen Akteuren, die strukturellen Beziehungen von Objekten der Wissensrepräsentation (z.B. von wie Artefakten) und der Formalisierung von Wissen und schließlich um Beziehungen zwischen strukturellen Elementen der Wissensorganisation, wie sie durch mentale Modelle oder Frames beschrieben werden können.[9].

Substrukturen dieser Netzwerke, wie die Beziehungen von Akteuren oder auch Netzwerke, die Austauschprozesse beschreiben, sind hierbei schon seit langem Gegenstand der Netzwerkforschung in der Soziologie und den Wirtschaftswissenschaften. Mit der historischen Netzwerkforschung wird seit den 2010er Jahren die soziale Netzwerkforschung als historisches Instrument eingesetzt.

Aufbauend auf vorhandenen Daten ist es häufig relativ einfach, Netzwerkstrukturen aufzubauen und Beziehungs- und Einflussnetzwerke zu visualisieren. Um den Apparat der mathematischen Graphentheorie auf diese Netzwerke anzuwenden, ist jedoch eine genaue, nachvollziehbare Klassifizierung der einzelnen Knoten und Kanten dieser Netzwerke unerlässlich. Hier ist das Zusammenführen der semantischen Modellierung und der Netzwerkanalyse vielversprechend.

Die im Folgenden vorgestellten Ansätze und Methoden sollen und können nicht die detaillierte historische Fallstudie ersetzen. Die neuen Methoden sind ein weiterer Baustein im Werkzeugkasten der Historikerin und des Historikers. Sie dienen als heuristische Instrumente und müssen wie jede andere Methodik unter historisch-kritischem Blickwinkel und als Teil eines hermeneutischen Prozesses gesehen werden. Wie das Auswählen, die Transkription und Übersetzung von Quellen stellt das Erstellen von Datenstrukturen, die maschinell auswertbare Kodierung von Informationen und die Auswahl von Algorithmen eine Interpretationsleistung dar und ersetzen diese nicht.

Die vorliegende Arbeit hat zum Ziel, Methoden und Verfahren vorzustellen, die in diesem Sinne von Historikern genutzt werden können, um Quellen systematisch computergestützt zu analysieren. Es wird insbesondere beleuchtet, welche Stellung diese Methoden im historischen Diskurs haben und wie umgekehrt die Interaktion zwischen informationstechnischer und historischer Sichtweise sich beeinflussen.

Insbesondere geht es um die Frage, wie ein interdisziplinärer theoretischer Rahmen aussieht, in dem Wissensorganisation und Wissensdynamik beschrieben und untersucht werden können.

1.1.3 Praktische Herausforderungen

Neben den theoretischen Herausforderungen sind Arbeitsumgebungen und Methoden zu entwickeln, die von den Fachwissenschaftlern eingesetzt werden können. In der täglichen Arbeit muss dabei ein Kompromiss zwischen Benutzerfreundlichkeit und Flexibilität gefunden werden. Außerdem müssen Ergebnisse nachhaltig verfügbar gehalten werden können. Zugleich ist ein modularer Ansatz sowohl aus ökonomischen wie aus inhaltlichen Gründen notwendig, der unterschiedlichste Anwendungen möglichst nahtlos zusammenführen kann. Von der Nutzerseite hat sich in langen gemeinsamen Diskussionen von Entwicklern und Fachwissenschaftlern gezeigt, dass eine auf einem elektronischen

[9] Zur Rolle von mentalen Modellen in diesem Kontext siehe [93, S.38ff] sowie F-Logic [255, 7].

Notizbuch aufbauende Lösung ein vielversprechender Ansatz sein könnte [36], wie er etwa innerhalb von Mathematica [170] umgesetzt ist. Da zumindest in den Kernfunktionalitäten auf OpenSource-Lösungen gesetzt werden soll, fiel die Entscheidung in unserem konkreten Fall auf Jupyter bzw. iPython [127] als Kern für eine interaktive Arbeitsumgebung, die nun in gemeinsamer Arbeit von MPIWG und der Digital Innovation Group der Arizona State University entwickelt wird. Die Grundstruktur lässt sich folgendermaßen zusammenfassen:

- jupyterbasierte Notizbücher,
- Pythonbibliotheken für die Analyse der Daten, die in die Notizbücher eingebunden werden können und einen einfachen Zugriff auf die Daten ermöglichen,
- ein über Webfrontends und REST-Bibliotheken ansprechbarer Server zur Publikation von Auswertungsergebnissen,
- die Anbindung eines Triplestores sowie einer Graphdatenbank,
- Prozesse, die längere Rechenzeiten benötigen, werden in externe Services ausgelagert.

1.2 Graphen und Netzwerke: Eine Begriffsklärung

Die terminologische Unterscheidung zwischen Graphen und Netzwerken ist nicht immer eindeutig. Den Terminus *Graph* benutze ich weitestgehend im Kontext der mathematischen Graphentheorie sowie im Kontext der semantischen Modellierung. Beiden ist gemeinsam, dass hier Strukturen über Ontologien oder über einen mathematischen Formalismus klar definierte Bedeutungen zugeordnet werden. *Netzwerk* benutze ich in der Regel im Kontext von sozialen Netzwerken und allgemein für Systeme, in denen Austauschbeziehungen zwischen den sie konstituierenden Teilen bestehen. Der Begriff ist daher dem umgangssprachlichen Gebrauch von Netzwerk nahe, während Graphen immer eine mathematische Repräsentation besitzen. Es lassen sich jedoch nicht immer Kontexte vermeiden, in denen die Begriffe weitestgehend synonym gebraucht werden. Im Allgemeinen sollte dieses nicht zu Verwirrung führen.[10]

1.3 Digitale Geisteswissenschaften oder digitale Wissenschaft?

Bevor ich im letzten Abschnitt dieser Einführung auf die konkrete Struktur dieser Arbeit eingehen werde, möchte ich doch schon zu Beginn mit einigen wenigen einführenden Bemerkungen auf die Funktion digitaler Methoden in der wissenschaftlichen Forschung eingehen. Diese Überlegungen sind einerseits Motivation für die aus meiner Sicht auch in der historischen Forschung notwendige Beschäftigung mit digitalen Methoden, andererseits greifen sie auch einige wenige der Argumente auf, mit denen die Skepsis gegenüber diesen Methoden häufig begründet wird.

[10] Die Probleme zeigen sich schon bei der Terminologie, die in den Softwarepaketen genutzt wird, die in der Regel zum Einsatz kommen. Die beiden weit verbreiteten Pakete zur Netzwerkanalyse in Python sind *igraph* und *networkx*. Bei beiden ist die Repräsentation eines Netzwerkes jeweils durch ein Objekt der Klasse *graph* gegeben.

In den Geisteswissenschaften wird – häufig auch zu Recht – viel über die Rolle von rechnergestützten Methoden gestritten. Die Entwicklung der terrestrischen Navigation mit Hilfe des Stands der Sterne, mit Wegmarken wie Pflanzen, Gebirgen und anderen unmittelbaren Sinneseindrücken, dann per Kompass, mit immer ausgereifteren Kartenwerken bis hin zur Satellitennavigation, kann hier als Parallele dienen – so problematisch Analogien auch immer sein mögen, vor allem wenn diese durch ihr gemeinsames Ziel – die Navigation durch die Natur – verlockend sind.

Die Navigation mit Hilfe von modernen Technologien macht die Erfahrung des Wanderns durch die Natur, des Entdeckens von neuen Wegen nicht obsolet. Sie ist eine neue Methodik, die teilweise Gleiches ermöglicht, um den Weg von A nach B zu finden, und teilweise Neues eröffnet, d. h. den schnelleren Weg oder alternative Wege, die man von selbst nicht beschreiten würde, da die Vorerfahrungen diese unter Umständen zu risikoreich erscheinen lassen, während neue Methoden jedoch eine Vorausberechnung ermöglichen und andere Wege plötzlich plausibler machen. Umgekehrt droht der Verlust von Erfahrungen und Eindrücken auf dem Weg, die zur Erkenntnisgewinnung genauso gehören wie das Erreichen eines konkreten geplanten Ziels. Diese Methoden ergänzen sich, sind Hilfsmittel auf dem Weg zum Verständnis der Organisation von Wissen, zur Erweiterung und zum Einsatz dieses Wissens für gesellschaftliche und technologische Veränderungen. Der Weg ins Archiv, die Einzelstudie, die materielle Erfahrung eines Objektes und vor allem der Prozess der Erkenntnis und Einordnung wird durch die digitalen Methoden nicht ersetzt. Denken müssen wir schon selbst, die „Schreib-mir-einen-Artikel-Taste“ ist weder in den Geistes- noch in den Naturwissenschaften das Ziel. Auch hier ist es hilfreich, dass uns der Begriff der Wissenschaft zur Verfügung steht, ohne in Sozial-, Geistes- und Naturwissenschaften unterscheiden zu müssen, wie es die begriffliche Trennung von *Science* und *Humanities* im angelsächsischen Sprachgebrauch notwendig macht.

Auch wenn der elektronische Navigator (zumeist) dabei hilft, ein Ziel zu finden, so entledigt er uns nicht der Aufgabe, dieses Ziel zu bestimmen. Er kann uns eventuell erreichbare Ziele aufzeigen, aber nicht das Ziel für uns auswählen. Und selbst hier versagt die Technik. Auch mit der ausgefeiltesten Ausrüstung finden wir immer wieder die *terra incognita*, die ohne das vorsichtige Vortasten, manchmal die Intuition, manchmal den Zufall nicht erschlossen werden kann, wie es jeder schon einmal erfahren hat, dem dank des Navigators die Sicht auf eine leider noch nicht gebaute Straße eröffnet wurde. Und so abgedroschen es klingen mag: auch der Weg ist das Ziel. Erfahrungen und Erkenntnisse auf dem Weg ersetzen nicht die Technik, dafür aber oft das verzweifelte Blättern im Atlas oder das Warten auf den Aufgang der Sonne oder das Aufklaren des Himmels, damit Wegzeichen wieder sichtbar werden.

Die Analogie hilft auch noch einen anderen Aspekt der „digitalen“ Geisteswissenschaften zu verstehen: die Notwendigkeit des Experiments mit den Methoden selbst. Dies ist für sich selbst Teil des Erkenntnisprozesses – ohne das Ausloten von Techniken und das Arbeiten mit Toy-Modellen sind neue Methoden nicht denkbar. Niemand (oder fast niemand) würde einen Zug auf die Reise schicken, der nicht im Vorhinein zumindest im Modell getestet wurde, bevor die ersten Passagiere zusteigen.

Der Zusatz „digital“ vor den Geisteswissenschaften ist ein Hinweis auf eine spezifische Methodik, ein Adjektiv zu „Geisteswissenschaften“, und das ist damit auch schon die erschöpfenden Beschreibung – auch wenn sich hier eine aus meiner Sicht faszinierende neue Methode auftut. Aber, um im Bild zu bleiben, ersetzt nichts einen Sommer in Wanderschuhen oder auf dem Fahrrad ohne Navi nur

mit der Karte gerüstet und manchmal gar ohne festes Ziel bei der Abfahrt – selbst das „Wann sind wir endlich da“ meines Sohnes ist dem nicht abträglich.

Damit ist der Rahmen für diese Arbeit gesetzt.

1.4 Übersicht über die Arbeit

Das Kapitel 2 beleuchtet das Verhältnis zwischen der Informatik und den Geschichtswissenschaften. Indem auf die Voraussetzungen des digitalen Arbeitens im spezifischen Bereich der Wissens- und Wissenschaftsgeschichte eingegangen wird, werden die Problem- und Fragestellungen erläutert, die sich daraus für die Realisierung von Arbeitsumgebungen und Datenstrukturen ergeben. Dazu ist eine Übersicht über die methodischen Grundlagen des wissenschaftshistorischen Ansatzes der historischen Epistemologie notwendig, auf dessen Grundlage der historisch-methodische Teil der Arbeit steht. Die in diesem Kontext bereits entwickelten und genutzten Methoden der digitalen Publikation von Forschungsergebnissen und -daten werden hier im Arbeits- und Entstehungskontext vorgestellt. Dies kann nur selektiv geschehen. Ziel ist es hierbei, einen breiten Überblick zu geben, der alle wesentlichen Ansätze aufgreift. Eine quantitative Vollständigkeit ist hier auf Grund der Vielzahl und Unterschiedlichkeit der Projekte nicht erreichbar.

In Kapitel 3 werden die Voraussetzungen für die Qualität und die Struktur der Daten für die historische Arbeit beleuchtet. Insbesondere wird der Frage nachgegangen, welche Anforderungen sich aus geisteswissenschaftlichen Forschungsfragen ergeben. Wesentliche Aspekte sind hier die Frage des Umgangs mit widersprüchlichen Daten, mit Autorschaft und Provenienz sowie das Ereigniskonzept als grundlegende Struktur für die Datenmodellierung historischer Prozesse [104, 18, 34].

Das Kapitel 4 geht auf die informationstechnischen Grundlagen der Datenmodellierung und des Schließens ein und beschäftigt sich mit den für das Verständnis des weiteren Verlaufs notwendigen grundlegenden technischen Fragen der Datenpublikation und Fragen im Bereich von Datenbanken. Dieses Kapitel dient insbesondere dazu, den weiteren Kontext vorzustellen, in dem diese Arbeit steht. Insofern beschäftigt es sich vertieft mit den für die Fallbeispiele notwendigen Methoden und kann Ansätze, die im Rahmen der Wissensrepräsentation ebenfalls eine wachsende Rolle spielen, wie zum Beispiel die Theorien der neuronalen Netze und evolutionäre Ansätze, nur anreißen.

Das Kapitel 5 stellt den Übergang von der Datenmodellierung zur Netzwerkforschung dar. Es führt in die für die Arbeit relevanten grundlegenden Begriffe der Netzwerkforschung ein. Auch hierzu existiert eine umfassende einführende Lehrbuchliteratur, auf die an den entsprechenden Stellen verwiesen wird. Im Zentrum steht eine genauere Definition der oben genannten unterschiedlichen Typen von historisch relevanten Netzwerken. Für die Interpretation mathematischer Charakteristiken besteht eine wesentliche Aufgabe bei der netzwerktheoretischen Beschreibung darin, ein klares historisches Verständnis der Funktion der einzelnen Bestandteile des Netzwerkes zu bekommen. Schließlich wird auf die von Claire Lemercier [140] thematisierte Gefahr der ahistorischen Interpretation von Netzwerken eingegangen.

Im Kapitel 6 werden die grundlegenden Verfahren und Methoden beschrieben, wie aus den in Datenbanken heterogen vorliegenden Daten die für die konkreten Studien notwendigen strukturierten

Daten erzeugt werden. Konkret werden Methoden zum Erzeugen von Strukturen in RDF und deren Speicherung und Ablage in einem Triplestore dargestellt, ebenso wird auf Probleme hingewiesen, die sich grundsätzlich bei den vorhandenen Datenstrukturen ergeben. Hier wird die konkrete Implementierung der Algorithmen auf Grundlage von Python mit einem Triplestore und einer Graphdatenbank vorgestellt. In der praktischen Arbeit hat es sich gezeigt, dass der Einsatz eines Triplestores ergänzt durch eine Graphdatenbank für die Analyse und Repräsentation von Wissensstrukturen sinnvoll ist. Das Verfahren zur Erstellung der für die Modellierung notwendigen benutzen Ontologie mittels Protegé [192] und die zur Umsetzung eingesetzten Pythonbibliotheken und interaktiven Notizbücher werden hier erläutert.

Die Kapitel 7, 8, 9 und 10, stellen die für das Projekt zentralen historischen Fallbeispiele vor und geben einen konkreten Überblick über die mit den neuen Methoden angegangen Fragestellungen und über die Probleme, die mit diesen Methoden gelöst werden können. Das erste Beispiel befasst sich hier zunächst mit klassischen Problemen der sozialen Netzwerkanalyse und der Zitationsanalysen. Die Ansätze werden hier am Beispiel des Projektes zur Geschichte der Allgemeinen Relativitätstheorie und eines Teilprojektes aus einem größeren Projekt zur Geschichte der Max-Planck-Gesellschaft geschildert.

Beiden Projekten ist die zentrale Fragestellung gemein, ob und wie sich die Etablierung von formalen Organisationsstrukturen in der Entwicklung von Netzwerkstrukturen widerspiegelt [25, 24]. Anhand dieser Beispiele wird der praktische Umgang mit komplexen und unvollständigen Datengrundlagen verdeutlicht. Es werden angepasste Tools und Algorithmen vorgestellt, die zur Bearbeitung der Daten angewandt wurden.

Der Schwerpunkt hier liegt einerseits auf den eigentlichen technischen Methoden, andererseits auf der Darstellung, wie sich diese Methoden aus dem Dialog mit den beteiligten Wissenschaftlern ergeben. Insbesondere wird hier auf die strukturellen Gemeinsamkeiten einer historisch-geisteswissenschaftlichen Beschreibung und einer ereignisbasierten Modellierung, wie sie im *Conceptual Reference Model* (*CRM*) der CIDOC [104] vorgenommen wurde, eingegangen. Ein wesentliches Ergebnis der gemeinsamen Arbeit innerhalb des Forschungsteams ist, dass sich dieser Modellierungsansatz leicht an die Denkweise historischer Forschung anpassen lässt und so als methodische Brücke zwischen den Disziplinen dienen kann.

Das Fallbeispiel zur Datenbank der administrativen Texte zum Bau der Kuppel des Doms in Florenz ist zunächst ein Beispiel dafür, wie der Übergang von einer strukturierten Datenbank zu einer Darstellung der Daten in RDF auf der Grundlage einer definierten Ontologie erfolgt. Es zeigt, die sich daraus ergebenen Möglichkeiten, Herausforderungen und Desiderate einer Anbindung an Daten in der Welt der *Linked Open Data*. Es wird dargelegt, wie sich aus der ontologiebasierten Darstellung RDF-Netzwerke erzeugen lassen, die sich mit Methoden der Netzwerktheorie analysieren lassen. Das Ergebnis ist eine deutliche verbesserte Einsicht in die Datengrundlage durch strukturierte Suchmöglichkeiten und einen verbesserten semantischen Zugang. Zugleich erlaubt die Studie einen Vergleich automatischer Klassifikationsmethoden mit manuellen Methoden der Klassifikation, die auf der Sachkenntnis der am Projekt beteiligten Wissenschaftlerinnen und Wissenschaftler beruhten.

Das von Matteo Valleriani am Max-Planck-Institut für Wissenschaftsgeschichte geleitete Projekt

zum Wissenssystem der *Sphaera des Sacrobosco* dient als Fallbeispiel, in dem Modellierung und Netzwerkanalyse direkt aufeinandertreffen. Langfristiges Ziel ist es hier, die geisteswissenschaftliche Frage nach der komplexen Struktur des Wissens über Astronomie, praktische Mathematik und Navigation, das in den verschiedenen Ausgaben der Sphaera seinen Niederschlag findet, in ein CRM-basiertes Modell zu übertragen. Hierbei werden zwei unterschiedliche Ebenen mit einander in Beziehung gesetzt. Es beleuchtet die Verflechtung von Wissensstrukturen und sozialen Strukturen als erste Ebene. Die Netzwerke, die hier untersucht werden, sind zunächst Netzwerke, deren Knoten selbst lokale *soziale Netzwerke* [219],[138] sind, die jeweils durch die Publikation einer neuen Edition der Sphaera identifiziert werden können und damit für die Ausbreitung und Modifikation des Wissens über die Sphaera stehen. Damit verbunden ist ein internes Netzwerk, das aus der inhaltlichen und formalen Struktur der Editionen aufbaut. Knoten sind hier auf der inhaltlichen Seite Themenfelder, die in den Editionen behandelt werden, und auf der formalen Seite die Struktur des Traktates, also die Gliederung in Kapitel und Unterkapitel etc.. Dieses doppelte Netzwerk bildet die zweite Ebene der Netzwerkstruktur des Wissens über die Sphaera.

Das abschließende Kapitel11 fasst die bisherigen Ergebnisse zusammen und zeigt die notwendigen nächsten Schritte im Hinblick auf die weiteren technischen Entwicklungen auf, um die in dieser Arbeit zusammengestellten Methoden zu einem integralen Bestandteil historisch-kritischer Forschungsarbeit zu machen.

1.5 Konventionen im Druck

Im Folgenden benutzen wir Fettdruck für Klassen und Eigenschaften, diese sind im Anhang 14 aufgeführt. XML-Tags werden als <*tag*> geschrieben. Dateinamen sind kursiv gedruckt. Pythonnotizbücher enden auf .pynb und sind als *foo.pynb* gekennzeichnet. Im Anhang 13 sind diese noch einmal alphabetisch aufgeführt, dort finden sich auch Hinweise, wie auf diese zugegriffen werden kann. Termini, die sich im Termindex (Anhang 14) wiederfinden, sind in dieser Form *kursiv* ausgezeichnet. Diese Ausdrücke sind nicht im gesamten Text kursiv, sondern nur bei ihrer ersten Erwähnung oder wenn es besonders wichtig ist, darauf hinzuweisen, dass es sich hier um einen technischen Ausdruck handelt. URL von Graphen und Entities bezeichnen wir als *http://mygraph*. Auch diese sind im Anhang verzeichnet.

1.6 Danke!

Wie – hoffentlich – schon diese Einleitung gezeigt hat, ist der im Folgenden dargestellte Ansatz ohne interdisziplinäre Zusammenarbeit nicht denkbar. Diskussion und Kooperation sind die Grundlage einer computergestützten Geisteswissenschaft. Über die lange Zeit, in der die Arbeit gewachsen ist, bin ich daher vielen Kolleginnen und Kollegen, Freundinnen und Freunden zu Dank verpflichtet, die ich hier unmöglich alle aufzählen kann. Daher also zunächst eine Entschuldigung an alle, die ich im Folgenden nicht nenne. Ich hoffe, Ihr wisst alle, wie dankbar ich Euch für Eure Unterstützung bin. Ohne die enge Zusammenarbeit in und mit der Abteilung 1 des Max-Planck-Institutes für Wissen-

schaftsgeschichte wäre keines der geschilderten Projekte möglich gewesen, mein Dank gilt dafür allen Kolleginnen und Kollegen dort, besonders Jürgen Renn, Matteo Valleriani, Matthias Schemmel und Jochen Büttner, sowie dem langjährigen Leiter der Bibliothek des MPIWG, Urs Schoepflin. Anregungen in den letzten Jahren gingen besonders auch von den Mitarbeiterinnen und Mitarbeitern des Projektes zur Geschichte der Max-Planck-Gesellschaft aus, auch dafür hier vielen Dank an Euch alle. Auf Roberto Lallis und Felix Langes Beiträge werde ich später noch besonders eingehen. Julia Damerow, Gerd Graßhoff und Manfred Laubichler standen immer mit Rat und Tat zur Seite, wenn es nötig war - und das war oft. Bei Robert Casties bedanke ich mich besonders für seine freundschaftliche Zusammenarbeit seit unseren gemeinsamen Tagen beim Aufbau der IT-Gruppe des MPIWG. Mein Dank gilt den Direktorinnen und Direktoren des MPIWG für Ihre Unterstützung.

Allen DARIAHnerinnen und DARIAHnern möchte ich ausdrücklichen danken, ihnen verdanke ich die Möglichkeit, über Disziplinengrenzen hinweg offen auch unausgegorene Ideen diskutieren zu können.

Günther Görz danke ich dafür, dass er nie die Hoffnung aufgegeben hat, dass diese Arbeit einmal abgeschlossen werden wird und vor allem natürlich für seinen Rat und die vielen Diskussionen um die Deutung von CRM und seiner Anwendung über die lange Zeit, in der diese Arbeit entstanden ist. Klaus Meyer-Wegener gilt mein Dank, dass er sich bereit erklärt hat, sich dieses Themas anzunehmen, und für seine wertvollen Hinweise.

Zwei Kollegen, Peter Damerow und Malcolm Hyman, sind leider viel zu früh von uns gegangen, ohne sie wäre vieles von dem, was ich im Folgenden beschreiben werden, niemals aus seinen Kinderschuhen herausgewachsen.

Schließlich soll Susanne Grandel, die den Kampf gegen meine nicht immer überzeugende Orthographie aufgenommen hat und die Arbeit Korrektur gelesen hat, nicht unerwähnt bleiben.

Und natürlich gilt mein Dank meiner Frau Lucy Norris und meinem Sohn Florian, die die Jahre mit mir durchgehalten haben und ohne deren – nicht immer nur leichten – Druck die vielen Enden dieser Arbeit, die über Jahre gewachsen sind, wahrscheinlich nie zusammengefügt worden wären.

Teil I

Grundlagen

Die folgenden einführenden Bemerkungen sollen einen Einblick in die Methoden der Geisteswissenschaften und der Informatik geben, die notwendig sind, um die anschließenden praktischen Teile einordnen zu können. Dieses kann weder eine vollständige Einführung in die historische Methode noch eine Einführung in die Informatik sein. Zwangsläufig sind Zuspitzungen und Kürzungen notwendig. Vor allem verfolgt dieser Teil nicht den Anspruch einer wissenschaftstheoretischen Grundlegung geisteswissenschaftlicher Arbeitsweisen. Vielmehr geht es um eine pragmatische Beschreibung und Einordnung von Arbeitsschritten im geisteswissenschaftliche Forschungsprozess, um den Austausch zwischen Informatik und Geisteswissenschaften über Verfahren und Methoden zu vereinfachen. Forschung und angewandte Methoden sind auch in den Geisteswissenschaften abhängig von gesellschaftlichen Kontexten und technologischen Möglichkeiten. Ersteres gehört zum Selbstverständnis der Geisteswissenschaften. Letzteres wird nur zu gerne vergessen. Auch wenn sich die Beispiele in den praktischen Kapitel auf sehr konkrete Projekte und Probleme beziehen, ist es gerade dieser Kontext, der die tiefere Beschäftigung mit digitalen Methoden notwendig macht. Insbesondere sind neue Methoden und neue Publikationspraktiken und -verfahren untrennbar verbunden. In diesem Sinne geht dieses Kapitel auf die Bedeutung von Information und Informationsinfrastrukturen im Netzwerkzeitalter ein, in die sich geisteswissenschaftliche Ergebnisse einbetten lassen müssen, wobei hier aus meiner Sicht Open Source und Open Science eine zentrale Rolle spielen.

Um die Rolle dieser Methoden im Forschungszyklus der Wissenschaftsgeschichte nachvollziehbar zu machen, gibt dieses Kapitel eine kurze Einführung in die historische Epistemologie und die Konsequenzen, die sich aus diesem Ansatz konkret im Hinblick auf die Auswahl von digitaler Methoden ergeben.

Kapitel 2

Fragestellungen der Geisteswissenschaften als Herausforderung an die Informatik

Schon seit langem sind digitale Methoden ein fester Bestandteil der wissenschaftshistorischen Forschung. Die ersten Ansätze gehen weit hinter die Erfindung und Durchsetzung des World-Wide-Web (WWW) zurück. Geisteswissenschaftliche Projekte mit dem Ziel, das neue Medium als Forschungsinstrument und neue Publikationsplattform zu nutzen, sind so alt wie das Internet. Eine Geschichte der Geisteswissenschaft im Internet wäre Thema für eine eigene Arbeit und würde den Rahmen hier sprengen. Doch ist es vor allem für Leser aus der Informatik nützlich, eine Vorstellung von den Herausforderungen zu bekommen, die sich aus der Praxis in der historischen Forschung ergeben.

Die Erfahrungen und Lösungsansätze, die sich aus langjähriger Arbeit an Projekten am Max-Planck-Institut für Wissenschaftsgeschichte speisen, sollen daher einen Einblick in diesen Mikrokosmos der praktischen Informatik geben.

Um die unterschiedlichen Projektansätze besser zu verstehen, lohnt es sich, drei eng verwobene aber tendenziell unterschiedliche Stränge des digitalen Arbeitens zu unterscheiden: die Aufbereitung und Verfügbarmachung von Quellen, die Anwendung computergestützter Methoden und von Algorithmen zur Analyse von Quellen und zum Test von Hypothesen und schließlich die Publikation von Ergebnissen und Daten. Alle diese Stränge sind in unterschiedlichem Maße Teil dieser Arbeit, wobei ein deutlicher Schwerpunkt auf der Verbindung des ersten und zweiten Aspektes liegt, jedoch der Publikationsaspekt als zentraler Teil jeder wissenschaftlicher Arbeit nicht außer Acht gelassen werden kann.

2.1 Symmetrien und Asymmetrien – Zu einer Epistemologie des Internet

Die mit der Digitalisierung verbundenen grundsätzlichen strukturellen gesellschaftlichen Veränderungen wurden bisher weitestgehend auf die Veränderung der Kommunikationswege, Produktionsverfah-

ren und sozialen Interaktion reduziert. Die fundamentalen Auswirkungen auf Wissenssysteme wurden dabei vernachlässigt. Nach der Digitalisierung der Kommunikationswege durch das Internet und der voranschreitenden Digitalisierung sozialer Beziehungen durch die sozialen Medien stehen wir nun vor der Herausforderung der Digitalisierung des Wissenssystems selbst. Strukturell bedeutete Digitalisierung zugleich Virtualisierung und die Stärkung von Netzwerkstrukturen als Informationsträger. Nach der Virtualisierung der Kommunikation durch das World-Wide-Web und sozialer Strukturen durch die sozialen Medien beobachten wir nun eine Virtualisierung der Wissensstrukturen selbst – und damit gewissermaßen eine Virtualisierung der Fakten. Die viel diskutierten „alternativen“ Fakten der Trump-Präsidentschaft sind hier nur ein – wenn auch besonders plakativer – Teil dieses Prozesses. Die mit der Virtualisierung verbundene Loslösung der Internetwelt von der physischen Außenwelt wird zwar seit einiger Zeit im Hinblick auf die sozialen Medien kritisch auch aus wissenschaftlicher Sicht betrachtet, Wissenschaftler (und auch Politiker) sehen in der Nicht-Akzeptanz ihrer wissenschaftlichen Forschungsergebnisse im Netz jedoch immer noch eher ein Kommunikationsproblem anstatt eines tiefgreifenden Problems der Wissensrepräsentation im Netz selbst. Dahinter steht eine Mißkonzeption, die das in den Strukturen des Internet repräsentierte Wissen auf der einen Seite und die physische Außenwelt auf der anderen Seite als parallele Systeme verstehen, die sich lediglich im punktuellen Austausch befinden.

Die Gefahr eines neuen Relativismus ist dabei evident und findet in der gegenwärtigen politischen Diskussion ihren dramatischen Ausdruck. Vor diesem Hintergrund ist die Krise nicht eine Krise der Kommunikation von Fakten, sondern Ausdruck einer grundlegenden Krise des Begriffs von Evidenz im Netzzeitalter. Nötig ist ein epistemologischer Ansatz, der reale – physisch erfahrbare – und virtuelle Knotenpunkte in enge Beziehung setzt und sie als Teil *eines* Netzes sieht. Notwendig für Orientierung in diesem komplexen System ist die Kompetenz, die unterschiedliche Evidenz und Funktion dieser Knoten zu bewerten. Benötigt werden Strukturen der Rückbindung des virtuellen Netzes an physische Realität, dafür bedarf es physischer Orte und Strukturen.

2.1.1 Informationsrevolutionen und -infrastrukturen

Informationsrevolutionen verbunden mit dem Aufbau neuer Informationsinfrastrukturen sind kein Spezifikum der Internetrevolution des 20. und 21. Jahrhundert. Die Druckerpresse, der Aufbau des Postsystems, die Erfindung der Telegraphie, das Aufkommen von Radio und Fernsehen haben stets zu grundlegenden Umwälzungen des Kommunikationsverhaltens und damit auch zu gesellschaftlichen Veränderungen geführt. Hierbei gibt es grundlegende Parallelen. Zumindest prinzipiell hat sich mit den neuen Kommunikationsformen das Verhältnis zwischen Sendern und Empfängern von Informationen verschoben. Häufig führten diese auch zunächst zu einer Öffnung von Informationskanälen zwischen breiteren gesellschaftlichen Schichten. Dieses gilt sowohl für die Sender als auch für die Empfänger: Kommunikation wurde zunächst erleichtert und damit auch demokratischer. Zugleich findet sich ein paralleler Trend im Versuch von staatlicher Seite und von privater Seite, Kontrolle über die Kommunikationswege zu übernehmen. Man beobachtet – wenn auch auf sehr unterschiedlichen Zeitskalen – die Kommerzialisierung und Monopolisierung der entsprechenden Medien und damit auch eine Privatisierung der Profite, von der Thurn-und-Taxischen Post über die großen Verlagshäuser bis hin zu

Google und Facebook. Und noch etwas ist auffällig: die Reduktion des Infrastrukturgedankens auf die technischen Funktionen, d.h. den Straßen- und Wegebau für die Post, Leitungswege und Sendeanlagen für die Kommunikation und nun die technischen Leitungen zu den Endkunden. Diese Tendenzen wirken den anfänglichen mit den neuen Technologien verbundenen Hoffnungen auf eine steigende Symmetrie von Sendern und Empfängern von Informationen entgegen.

Die Umwälzung von Kommunikationswegen ist stets mehr als eine quantitative Ausweitung, sie bedeutet immer auch eine grundlegende Veränderung der damit verbundenen notwendigen Kulturtechniken und es bedarf einer Koevolution der Bildungswege und der gesellschaftlichen Strukturen. In diesem Sinne ist nicht Wissen an sich Macht, sondern das Wissen um die Kommunikationswege und -strukturen.

Doch wie jede Technologierevolution hat das Netz zugleich einen grundlegend neuen Ansatz in die Wissensökonomie gebracht. Wissen hat sich delokalisiert, Wissen im Netz ist durch das gesamte Netz als Einheit repräsentiert und nicht durch seine Komponenten, seine Kanten und Knoten.[1] Dieses bedeutet jedoch keine Dematerialisierung von Wissen – ein Eindruck, der durch Sprachbilder wie der Cloud vorgetäuscht wird, und es bedeutet nicht notwendig Redundanz oder Demokratisierung von Wissen. Das Netz als Struktur kann Redundanz ermöglichen, in dem der Ausfall von Knoten nicht relevant ist und in dem keine Hoheit über Informationen besteht. Redundanz ist aber nicht zwingend im Netz. Der Gründungsmythos des Internet ist mit dieser Vorstellung verbunden, die Realität zeigt doch eher in die andere Richtung, in der Ausfälle von einzelnen Knoten zwar kompensiert werden, so dass die Transportwege sicher bleiben. Informationen selbst sind jedoch häufig mit dem Ausfall von Knoten nicht mehr zugänglich.

Zugleich ist das Bild der Cloud auch ein gerade von den Internetmonopolisten gepflegtes Bild. Vermittelt wird der Eindruck der Fluidität und Immaterialität des Wissens. Mit Begriffen der „Green-IT“ ist zwar die ökologische Implikation auch der Wissensökonomie durch ihren Ressourcenverbrauch in die öffentliche Diskussion geraten. Dieses impliziert in der öffentlichen Debatte bisher jedoch nicht die Verbindung von Materialität von Information und der damit verbundenen Frage des Besitzes von Informationen auch im Sinne ihrer physischen Existenz. Zwar ist dieses Teil einer Expertendebatte, aber sie bestimmt nicht das Nutzerverhalten.

2.1.2 Internet der Dinge und nicht des Wissens

Damit verbunden ist eine weitere Tendenz. Die Zukunftsdiskussion um das Internet ist geprägt vom Internet der Dinge – Kühlschränke, Autos und Spielzeuge, die miteinander kommunizieren und unsere Wünsche erfüllen. Im Hintergrund dieser Diskussion bleibt der eigentliche Gegenstand dieser Kommunikation: das Wissen über die Konsumenten, über ihre Angewohnheiten und Verhaltensmuster. Die mit Symmetrie von Sender und Empfänger von Information erweckten Hoffnungen erfüllen sich nun eher in einer Form einer Dystopie der Kontrollierbarkeit anstelle der Utopie der Freiheit.

[1] In Abschnitt 2.1 sind wir schon darauf eingegangen, dass dieses auch Schattenseiten hat. Auch bewusste Falschmeldungen und Unwahrheiten finden sich eingebettet in das Wissenssystem “Netz“ wieder. Methoden müssen gefunden werden, um diese zu isolieren und bewusst werden zu lassen.

Dem entgegen steht das Zukunftsbild des epistemischen Web, das das semantische Netz ablöst und aufhebt. Indem der Browser zum Interagenten wird, wird der Nutzer vom Konsumenten zum Agenten. Hier ist die Cloud auch eine Chance. In noch vor fünf Jahren nicht gedachtem Maße können mit netzbasierten Microservices und dank HTML5 Anwendungen und Daten kombiniert und angereichert werden. Hier stehen wir noch am Scheideweg zwischen mehr Autonomie der Nutzer und dem Ausliefern von Informationen an Dritte.

Bei allem Mitgefühl für Don Quichote und der Sympathie seines Kampfes gegen die Windmühlen, die Entwicklung des Netzes anzuhalten, wird keine Lösung sein. Der Weg hin zu einer Wissensgesellschaft mit selbstbestimmten Agenten steht immer noch offen. Die Speicherung von Daten über unser Verhalten wird sich nicht mehr aufhalten lassen. Es geht nun darum, Gestaltungsspielräume aufzutun, um eine Governance des Internets und damit der Wissensgesellschaft zu ermöglichen.

Voraussetzungen dafür sind vor allem: die Offenlegung der Mechanismen, die zur Verknüpfung von Informationen führen und die Schaffung von Räumen des freien unmittelbaren Austausches. Daraus ergibt sich unmittelbar die öffentliche Aufgabe des Ausbaus und Umbaus der Bibliotheken und Museen, die Unterstützung der Kulturlandschaft und schließlich vor allem auch der Bildungseinrichtungen. Die Diskussion um alternativen Fakten und Parallelwelten greifen zu kurz. Akteure in einer Parallelwelt könnten wir amüsiert verfolgen (oder auch nicht, da sie sich unserer Erkenntnis entziehen würden), doch die Krise ist eine andere. Die Krise ist auch und vor allem eine Krise der vernetzten Welt. Es ist das Symbol einer Entwicklung, die die virtuelle Welt von der Außenwelt entkoppelt, in der sich abgeschlossene Informationsnetze entwickeln können, die Wissen nicht mehr an die Materialität und physische Erfahrung rückbinden. Es ist eine Welt, in der alle Knoten eines Netzes gleichwertig werden und damit eine Krise dessen, was Evidenz bedeutet. Je nach Zielgruppe können so Teilnetze aktiviert werden, die in sich weitgehend konsistent sind. So können widersprüchliche Aussagen gleichzeitig existieren, da die Teilnetze nie oder nur in schwachen Verbindungen überschritten werden. Der einzelne konstruiert in diesem Netz sein eigenes Wissenssystem und nimmt dabei nur zu gerne die Unterstützung anderer in Anspruch. Es entstehen isolierte oder nur lose gekoppelte Teilnetzwerke. Offensichtlich ist die Gefahr der Entstehung von Wissenssystemen, die nur sehr punktuell im Rahmen des eigenen Erfahrungshorizontes an die „Realwelt“ angebunden sind (Abbildungen 2.1 und 2.2)..[2]

[2]Eli Pariser führte in diesem Zusammenhang den Begriff der *Filterblasen* ein [182]. Entscheidend ist jedoch, dass der von mir geschilderte Prozess in eine grundlegende Veränderung des Wissenssystems eingebunden ist. Das WWW ist in diesem Prozess ohne Zweifel das dominierenden Medium der Informationsorganisation. Zugleich führen Globalisierung und die damit verbundene Veränderungen der Steuerungsmöglichkeiten durch die Nationalstaaten zu einem gefühlten oder auch tatsächlichen Rückgang der Gestaltungsmöglichkeiten des Einzelnen. Damit sind diese in gleicherweise relevant für das Bedürfnis der Erschaffung überschaubarer Einheiten. Inwiefern der Einfluss der Internetgiganten hier Ursache oder Wirkung ist, ist nicht ohne Zweifel zu bestimmen.

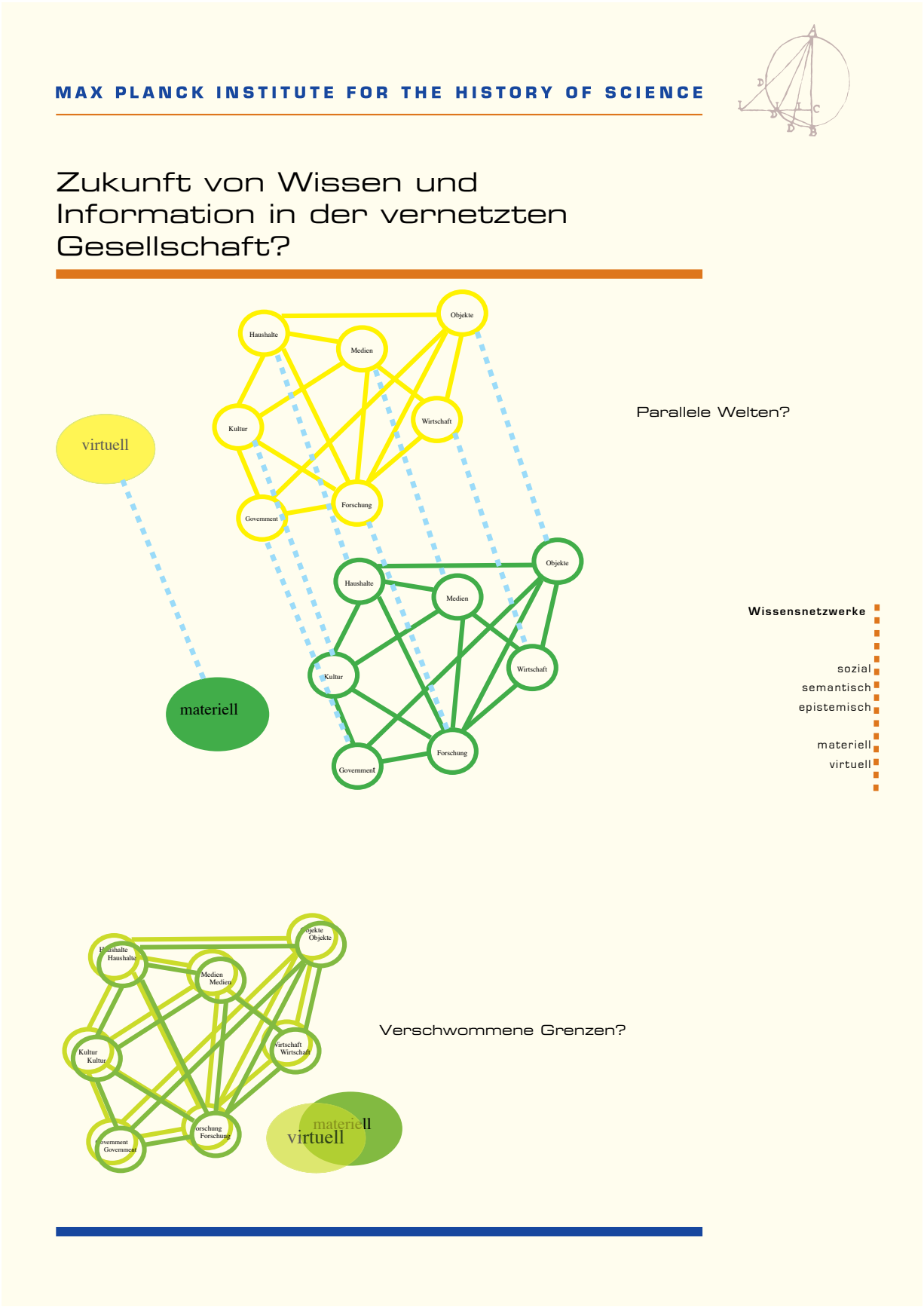


Abbildung 2.1: Die vernetzte Gesellschaft: Herausforderungen

Mit dem Antropozändiskurs und der möglichen Anerkennung des Anthropozän als ein neues erdgeschichtliches Zeitalter wird die Veränderung der Umwelt durch den Menschen als treibende Kraft der Umbrüche der Systems „Erde“ anerkannt. Die Analyse dieser Strukturen ist zwangsläufig unvollständig, wenn nicht auch die virtuellen Strukturen des Internet in diesen Diskurs einbezogen werden. Spätestens seit den frühen 2000er Jahren ist der Begriff des Netzwerkzeitalters in die Diskussion um die internationale Entwicklung eingegangen [59]. Jedoch stehen hierbei die technischen, sozialen und ökonomischen Auswirkungen im Zentrum. Die politische Diskussion um die Auswirkungen der Globalisierung nimmt die notwendige Reflexion über das Netz als globaler Ort nur am Rande auf [128]. Internetzensur und Abschalten von Zugängen zum Internet geben den falschen Eindruck, dass hinter dem Netzwerkzeitalter immer noch Strukturen bestehen, die mit den alten Mechanismen staatlicher lokaler Politik angegangen werden können – auch wenn es hier zumeist negative Beispiele staatlicher Regulierung sind. Hintergrund auch hier ist die Gleichsetzung des Netzwerkzeitalters mit dem Internetzeitalter und damit die Dominanz einer spezifischen Ausprägung des Netzwerks. Das Netzwerkzeitalter ist jedoch mehr. Es ist das komplexe System der globalen Handelsstrukturen virtuell und real, dem Austausch von Personen und Dienstleistungen, aber vor allem auch eine prinzipielle Veränderung, wie Wissen kommuniziert und strukturiert wird.

Das Netzwerkzeitalter ist zwar ohne Digitalität nicht denkbar. Diese ist jedoch lediglich die Voraussetzung, aber nicht die hinreichende Bedingung für das Netzwerkzeitalter.

2.1.3 Open Science: Open Data, Open Source und Open Information

Eine Konsequenz für die Wissensorganisation ist die Einheit von materiellen und virtuellen Knoten als Teil des gleichen Wissensnetzes. Die Realität des Netzes und die Realität der Außenwelt sind Bestandteil eines verflochtenen Systems.

Open Source, Open Data und Open Information sind hier keine Luxusdebatten, sondern eine notwendige Konsequenz aus dieser Erkenntnis. Wissenschaft kann sich vor dem Hintergrund den Luxus einer „closed“ Science nicht leisten, sondern Wissenschaft muss sich mit ihren Ergebnissen im Netz in der gleichen Selbstverständlichkeit bewegen wie die sozialen Akteure in Facebook und Twitter.

Dieses ist die Voraussetzung für die Rückkopplung von Netzrealität und Außenwelt. Offenheit, verbunden mit Kompetenzbildung, ist die notwendige Bedingung dafür, die Durchsetzung „alternativer Wahrheiten“ verhindern. Fakten müssen im Netz sichtbar sein, um sie in das Wissenssystem des Netzzeitalters integrieren zu können. Daraus ergibt sich eine zentrale Herausforderung an die Informatik, sie muss sich der Herausforderung stellen, die Akteure in die Lage zu versetzen, die Strukturen des Netzes zu verstehen. Ohne eine systematische Zusammenarbeit von Geistes- und Kulturwissenschaften sowie der Informatik ist diese Herausforderung nicht zu bewältigen.

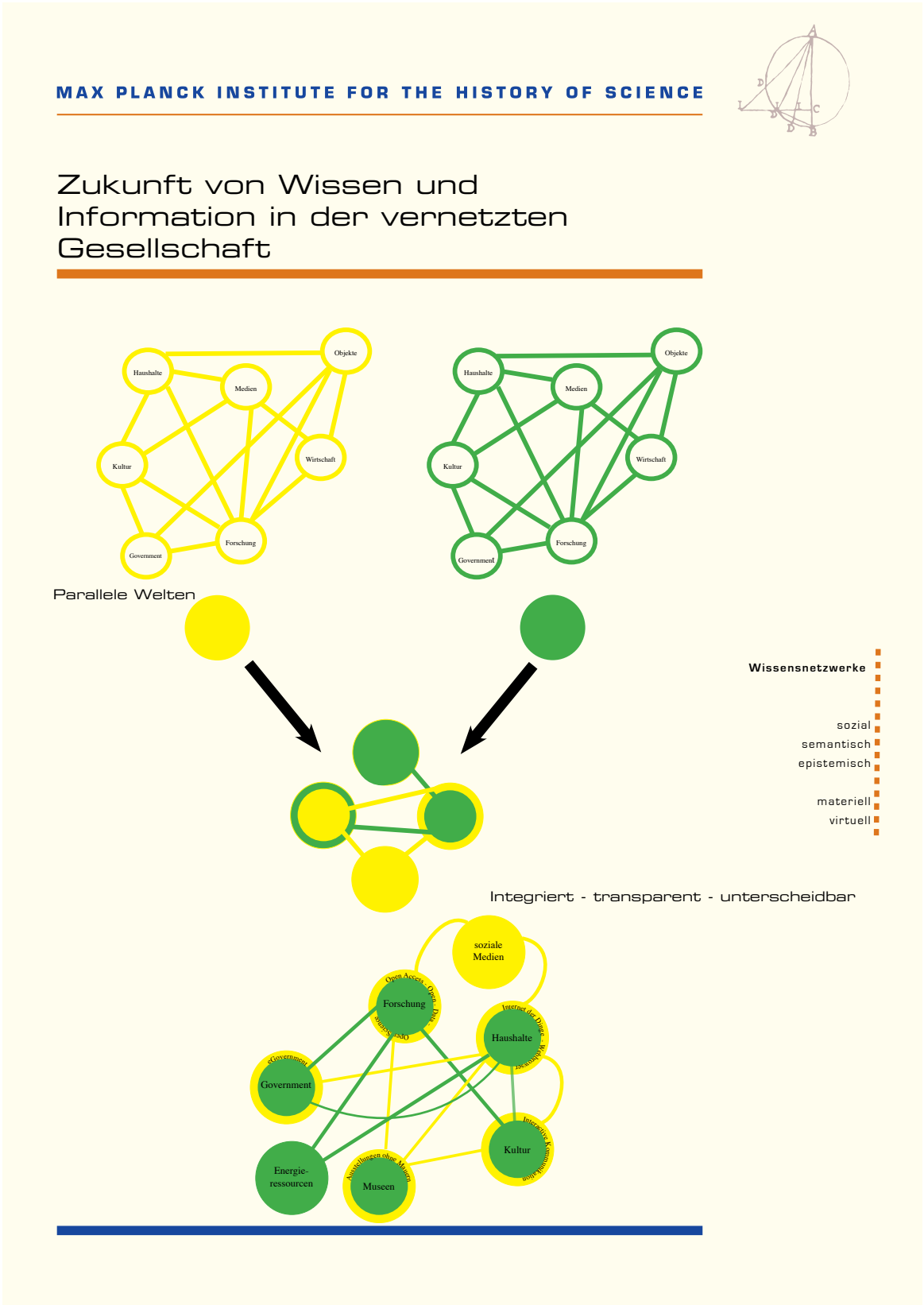


Abbildung 2.2: Die vernetzte Gesellschaft: Integration

2.2 Historische Epistemologie

Eine umfassende Einführung in die historische Epistemologie würde den Rahmen dieser Arbeit bei Weitem sprengen. Für eine ausführlichere Darstellung sei auf den Aufsatz von Jürgen Renn in [201] und das instruktive Büchlein von Hans-Jörg Rheinberger [208] verwiesen. Der Ansatz der historischen Epistemologie hat zum Ziel, „die einst sauber getrennten Kontexte der Rechtfertigung und der Entdeckung neuen Wissens wieder zusammenrücken.“ [208, S.11] Insbesondere sollen das Prozesshafte der Wissenschaft betont werden und die unterschiedlichen Faktoren, die Wissenschaft bedingen, im historischen Kontext analysiert werden; oder wie Jürgen Renn [201, S.241] es ausdrückt:

> Since the emergence of scientific disciplines is a process involving both social and cognitive factors, only an historical theory of scientific cognition which comprises both the social and the cognitive structures of science will be able to cope with the challenge to our understanding of science that is created by its growing interdisciplinary character. Such a theory, which I would like to call „historical epistemology“

Unter diesem Ansatz reicht es weder aus, sich bei der historischen Beschreibung von Entdeckungen auf die akribische Rekonstruktion des Prozesses zu konzentrieren, der zur Entdeckung geführt hat, noch kann die Geschichte der Wissenschaft als Geschichte der Individuen und Organisationen allein verstanden werden, die zum Prozess der Wissensentwicklung beigetragen haben. Es bedarf einer Analyse der komplexen Wechselbeziehungen zwischen Akteuren, der Entwicklung von Konzepten sowie der gesellschaftlichen und kognitiven Bedingungen, unter denen Wissenschaft stattfindet. Mit den Worten von Hans-Jörg Rheinberger definiert sich Epistemologie folgendermaßen [208, S.13]:

> [Der Begriff der Epistemologie] wird hier nicht einfach synonym für eine Theorie der Erkenntnis verwendet, die danach fragt, was Wissen zu wissenschaftlichem Wissen macht, wie dies für die klassische Tradition und insbesondere den angelsächsischen Sprachraum charakteristisch ist. Ich fasse unter dem Begriff der Epistemologie hier vielmehr, an den französischen Sprachgebrauch anknüpfend, die Reflexion auf die historischen Bedingungen, unter denen, und die Mittel, mit denen Dinge zu Objekten des Wissens gemacht werden, an denen der Prozess der wissenschaftlichen Erkenntnisgewinnung in Gang gesetzt sowie in Gang gehalten wird.

In Abschnitt 5.13 wird ausführlich dargestellt, wie sich aus diesem Ansatz eine netzwerk- und modellierungstheoretische Beschreibung ableiten lässt.

2.3 Digitale Projekte in den Geisteswissenschaften - Motivation

In den letzten Jahrzehnten haben digitale Methoden auch in die historisch arbeitenden Wissenschaften Einzug gehalten. Getrieben durch die Textwissenschaften verbunden mit den Anstrengungen von Bibliotheken ihre Bestände digital verfügbar zu machen, stehen textliche Quellen in vielen Bereichen für die wissenschaftliche Arbeit zur Verfügung.

2.3.1 Forschungsdatenzyklus und Forschungszyklus

Digitale Hilfsmittel unterstützen hierbei die Forschung in unterschiedlichen Phasen des wissenschaftlichen Forschungszyklus an sehr unterschiedlichen Stellen. Umfragen unter Geisteswissenschaftlern [227, 228] zeigen jedoch, dass nach wie vor die Unterstützung beim Schreiben von Artikeln und hier zumeist der Einsatz von Textverarbeitungsprogrammen sowie Hilfsmitteln zur Erstellung von Bibliographien den größten Anteil haben. Dann folgen Suchmaschinen und Datenbanken im Netz. Komplexere Methoden werden dagegen selten eingesetzt und im wesentlichen nur im Rahmen von einzelnen Projekten.

Um eine Übersicht über die bereits existierenden digitalen Methoden und ihre Zuordnung zu Forschungsphasen zu ermöglichen, wurden in den letzten Jahren Anstrengungen unternommen, den geisteswissenschaftlichen Forschungsprozess zu systematisieren. Genannt seien hier die geisteswissenschaftliche Forschungsontologie TaDiRAH [230] und das Projekt Bamboo [63] und die in diesem Kontext dort veröffentlichte systematische Liste von Tools. Grundsätzlich gehen die meisten dieser Ansätze auf die Scholarly Primitives von John Unsworth [245, 28] zurück. Abbildung 2.3 zeigt unterschiedliche Klassifkationsmodelle im Vergleich.

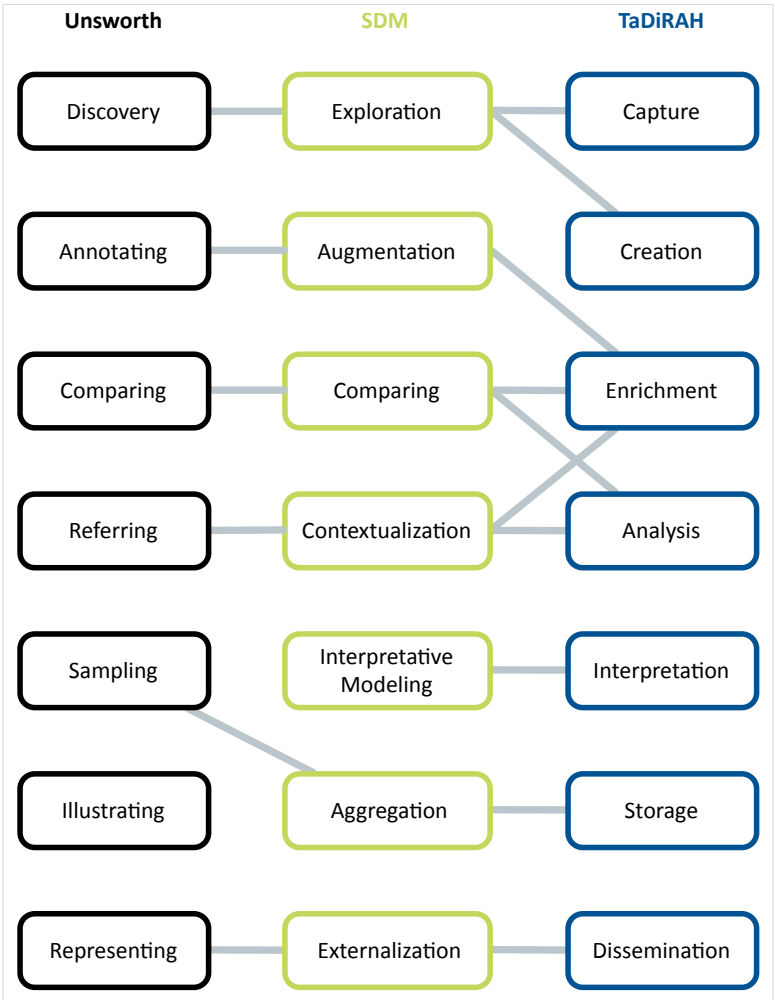


Abbildung 2.3: Scholarly Primitives [245] , TaDiRAH [230] und Scholarly Domain Model (SDM) [156] im Vergleich, Abbildung aus [139].

Der Forschungszyklus (Abb. 2.4), der im Rahmen von DARIAH-DE entwickelt wurde, versucht eine Konkretisierung der sehr abstrakten Vorschläge von Unsworth und der detaillierten Darstellung von TaDiRAH. Insbesondere soll der Zyklus sowohl für die Geisteswissenschaftler, die sich mit digitalen Methoden beschäftigen wollen, als auch für Entwickler von Tools für die Geisteswissenschaften verständlich sein. Dieser Zyklus dient hierbei als Verständigungsgrundlage im Sinne eines pragmatischen Hilfsmittels und hat nicht den Anspruch einer wissenschaftstheoretischen Analyse der Forschung.

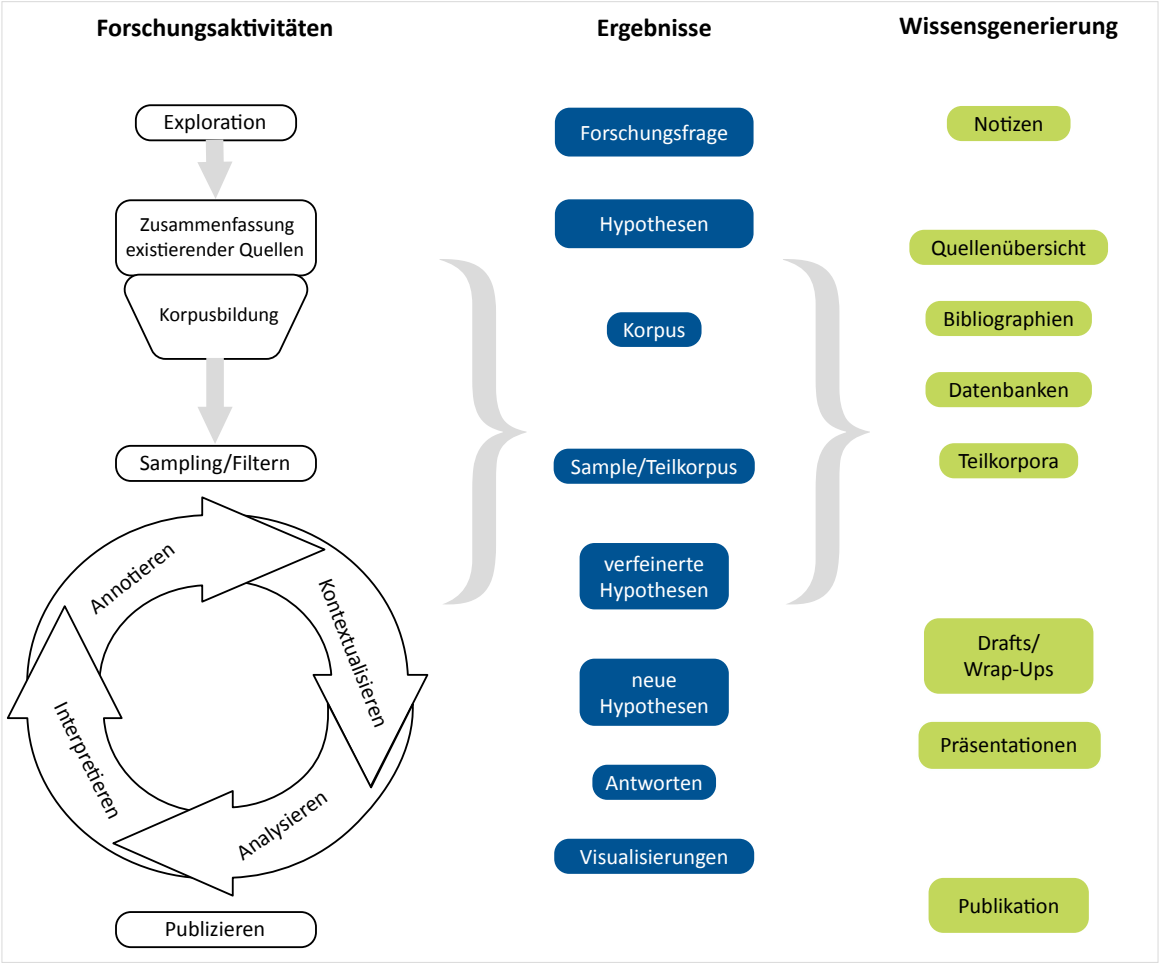


Abbildung 2.4: Forschungszyklus, Abbildung aus [139]

Die Spalten stellen hier die Hauptschritte beim Umgang mit Daten in der geisteswissenschaftlichen Forschung dar. Daten umfassen hier die in Abschnitt 2.5 dargestellte Breite verschiedener Formate und Strukturen. Da die Fallbeispiele alle im wesentlichen auf textuellen Daten beruhen, konzentriere ich mich im folgenden auf die Bearbeitung dieses Datentyps, wobei eine Übertragung auf Audiodateien, Bilder und Filme zwar eine technische Herausforderung darstellt, aber prinzipiell auf gleichen Arbeitsweisen beruht.

Parallel zum Forschungszyklus ergibt sich ein Zyklus von Forschungsdaten, die in den Wissenschaften anfallen (Abb. 2.5) . Auch hier dient der Zyklus in erster Linie dazu, eine Sensibilisierung für die Problematik zu erreichen und insbesondere den Datenbegriff auch bei Geisteswissenschaft-

lern zu etablieren. Ein Beispiel dafür wurde im Rahmen der AG Forschungsdaten von DARIAH-DE aufbauend auf dem Zyklus der Data Documentation Initiative (DDI) entwickelt [263].

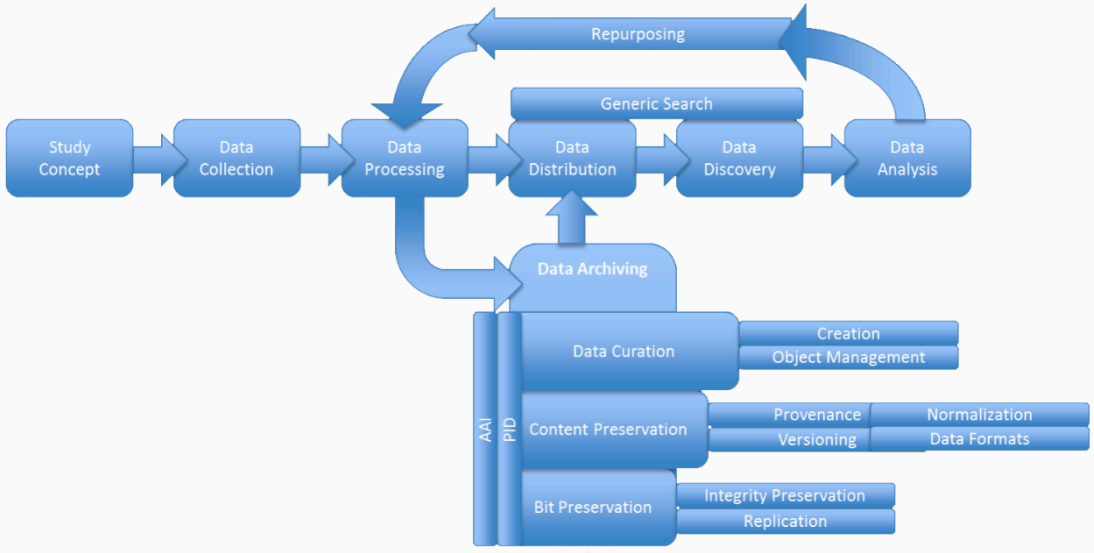


Abbildung 2.5: Forschungsdaten, übernommen von `https://www.textgrid.org/de_DE/web/old/bestehende-konzepte`

2.3.2 Geisteswissenschaftliche Prozeduren und Methoden der Informatik

Abbildung 2.6 stellt einen ersten Ansatz dar, geisteswissenschaftliche Prozeduren und Methoden der Informatik in einen Kontext zu setzen. Wie der Forschungskreislauf ist dieses Modell eine Orientierungshilfe zur Einordnung von Verfahren und Grundlage für eine weitere Diskussion. Das Modell ist das Ergebnis einer Einordnung auf Grund der Erfahrungen des Autors und basiert nicht auf einer quantitativen Studie. Es soll in erster Linie verdeutlichen, wie sich die folgenden Fallbeispiele in einen größeren Rahmen einbetten lassen, und soll den Blick auf Forschungsdesiderate und weitere Möglichkeiten der Kooperation zwischen Informatik und Geisteswissenschaften weiten.

Die dargestellte Reihenfolge gibt eine tentative Ordnung von Schritten an, die jedoch nicht als strikte Abfolge verstanden werden soll. In der Praxis ist die Ordnung sicher nicht immer genau diese und natürlich gibt es immer wieder Schritte vor und zurück, so dass sich Zyklen der Bearbeitung ergeben.

Die drei Spalten stellen drei grundlegende Prozesse der geisteswissenschaftlichen Forschung dar, bei denen computergestützte Methoden bereits heute zum Einsatz kommen, während in den Zeilen technische Methoden aufgezählt werden. Die Intensität der Färbung gibt die Intensität des Einsatzes dieser Methoden innerhalb der Prozeduren an. Die Anordnung ist so, dass sich von links nach rechts und von oben nach unten eine jeweils steigenden Komplexität ergibt; während die oberen Zeilen auf Methoden verweisen, die zwar unter Umständen noch einer technischen Entwicklung bedürfen, aber doch weitestgehend so generische Problemstellungen sind, dass die technische Entwicklung getrennt von der Arbeit mit diesen Methoden stark arbeitsteilig erfolgen kann, stellen die unteren Zeilen für beide Disziplinen, die Informationstechnologie und die Geisteswissenschaften, Herausforderungen

dar, die im interdisziplinären Kontext gelöst werden müssen. Schritte in diese Richtung sind zwar in Einzelfällen bereits gegangen worden, jedoch fehlt es an einer systematischen Kooperation. Ein Beispiel dafür ist, dass in der Informatik oftmals an einfach zugänglichen Korpora und Daten geforscht wird, wie etwa dem Twitterkorpus [70], Wikipedia [58], Facebook [145], Wordnet [275, 150] oder Daten aus dem Projekt Gutenberg [75].

Während umgekehrt geisteswissenschaftliche Institutionen auf großen Korpora und Datenmengen sitzen, die doch zu meist noch traditionell mittels manueller Durchsicht bearbeitet werden. Ein engeres Zusammenbringen würde Synergien auf beiden Seiten erzeugen. Die in dieser Arbeit vorgestellten Beispiele sollen eine Richtung aufzeigen, in die zukünftige kooperative Projekte gehen können. Insofern durchzieht das hier dargestellte Schema die Arbeit als ein roter Faden. Es wird an praktischen Beispielen dargestellt, wie sich geisteswissenschaftliche Forschung und Informatik gegenseitig bedingen.

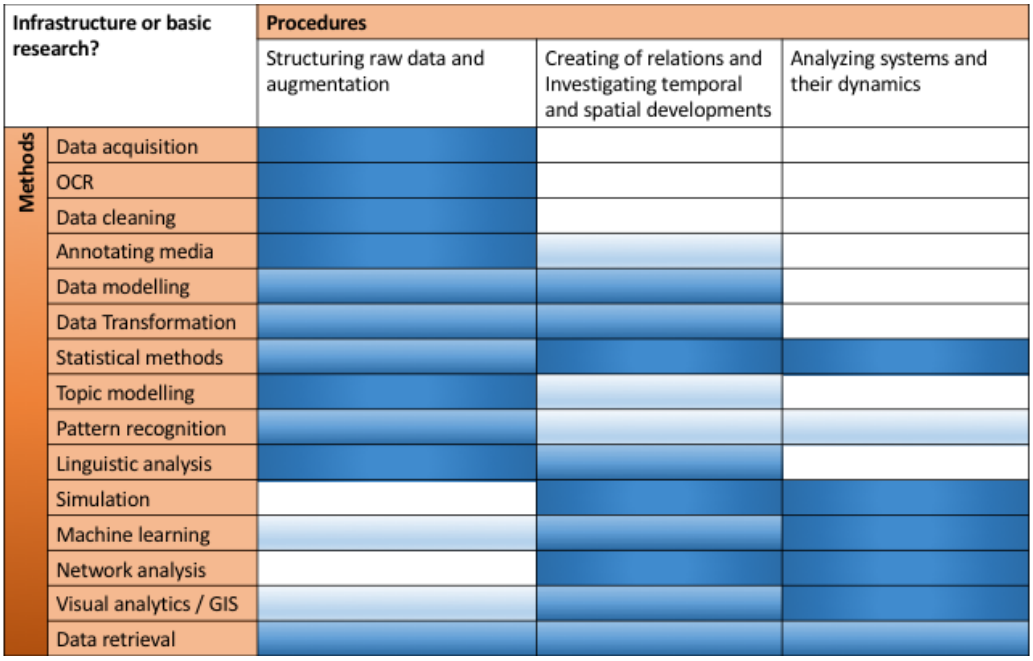


Abbildung 2.6: Prozeduren und Methoden in Geisteswissenschaften und Informatik: qualitativer Versuch einer Einordnung. Die Intensität der Färbung gibt die Intensität des Einsatzes dieser Methoden innerhalb der Prozeduren an.

2.4 Das weitere Umfeld - Digitale Quellen und Repositorien

Vor allem vorangetrieben durch die Editionswissenschaften und die Anstrengungen aus der Philologie und Linguistik stehen umfängliche Textkorpora zur Verfügung. Wissenschaftliche Editionsprojekte haben nun zumeist auch eine digitale Komponente oder werden immer mehr zu rein digitalen Projekten.[3]

Auch von Seiten der Museen werden verstärkt Anstrengungen unternommen, digitale Repräsentationen ihrer Objekte verfügbar zu machen. So weist die Europeana [74] mittlerweile einen großen

[3] Z.B. die Schriften von Isaak Newton in [2] und [236] oder die Korrespondenz von Charles Darwin [53].

Bestand von Museumsbeständen nach[4], auf die in Form von digitalen Faksimiles zugegriffen werden kann.[5] Nicht zuletzt – mit der Kunstgeschichte als ein treibender Motor – steigt die Anzahl von digitalen Projekten, die eine inhaltliche Erschließung bildlicher Daten ermöglichen [214]. Schließlich existieren umfangreiche Datenbankprojekte, die aufbereitete Daten für verschiedene Bereiche der historischen Forschung im Netz bereitstellen. Methoden im Kontext von Linked Open Data machen es in der Theorie möglich, diese Daten im Netz verfügbar zu machen und einer Nachnutzung zur Verfügung zu stellen.

Digital vorliegende Quellen sind auch in den historischen Geisteswissenschaften immer mehr zu einem zentralen Instrument für die wissenschaftliche Arbeit geworden. Datenbanken dienen hierbei als technisches Hilfsmittel im Rahmen eines Forschungsprojektes zur Organisation der Datenbestände, die während des Forschungsprozesses anfallen, als Nachweisinstrument für Quellen und werden zunehmend als Primärquellen selbst Gegenstand der historischen Forschung.

2.5 Digitale Quellen und Repositorien in der Wissenschaftsgeschichte

Am Max-Planck-Institut für Wissenschaftsgeschichte spielen digitale Publikationen und der digitale Zugriff auf Onlinebestände seit dessen Gründung 1994 eine zentrale Rolle. Die Wissenschaftsgeschichte als interdisziplinäre Forschungsrichtung ist auf Quellen aus unterschiedlichen Kontexten angewiesen. Sprach-, Kultur- und Disziplingrenzen müssen überwunden werden, um eine übergreifende Geschichte der Entstehung wissenschaftlichen Wissens nachzeichnen zu können. Hierbei geht es zum Einen um eine genaue Beobachtung und Analyse des Einzelfalles, zum andern um die Einbettung dieser Einzelfälle in den historischen Kontext der gegebenen Wissensbasis in der historischen Situation im Sinne einer historischen Epistemologie. Diese Kontexte historisch zu rekonstruieren, bedarf einer möglichst genauen Erfassung und Beschreibung der Quellen und Voraussetzungen, die im zeitgenössischen Kontext zur Verfügung standen. Eine virtuelle Rekonstruktion der historischen Situation mit Hilfe von elektronischen Arbeitsumgebungen bietet sich hier geradezu an. Welche Rolle Datenbanken in diesem Kontext spielen, ist ein wesentlicher Bestandteil dieser Arbeit. Beispiele aus der Arbeit des Institutes finden sich in so unterschiedlichen Bereichen wie der Cuneiform Digitial Library Initiative [40] oder dem Virtuellen Laboratorium der Physiologie [239].

Ein frühes Beispiel (1999), wie die Möglichkeiten von Hypertexten zur Publikation von Quellen benutzt werden können, findet sich in *Galileo Galilei's Notes on Motion* [85]. Hier wird ein vollständiges Manuskript Galileo Galileis im Internet einschließlich Transkriptionen und Nachzeichnungen zur Verfügung gestellt gemäß den damaligen technischen Möglichkeiten als statische Hypertexte in HTML. Gerade durch das Fehlen dynamisch generierter Elemente wird hierbei die Stärke einer Publikation als Netzwerk von Informationen deutlich. Dem Nutzer wird von den Autoren der Seite eine Interpretation und Deutung vorgeschlagen, die Einzelseiten bilden hierbei den interpretativen Kern. Diese bilden so ein Netzwerk von Informationen und Interpretation. Quelle und Deutung werden einerseits dank der Linkstrukturen in direkte Beziehung gesetzt, können jedoch auch getrennt betrachtet werden.

[4] Am 18.10.2017 waren es laut Webseite der Europeana 53.162.248 Datensätze.

[5] Zur Problematik siehe [226].

Anmerkungen und auch die Indizes schlagen eine Lesart der Quelle vor, zwingen sie aber nicht auf. Anders als in einer klassischen Publikation, die im wesentlichen einen Weg durch die Manuskripte fest vorgibt, bekommt der Rezipient mehrere Vorschläge für ein Narrativ durch die Referenzen der Seiten untereinander. Es sind mehrere Wege durch das Manuskript möglich, da eine Vielzahl von Verknüpfungen der Seiten untereinander besteht. Der Leser hat zudem das vollständige Manuskript vorliegen und kann seinen eigenen Weg finden.[6] Die statische Web-Site hat darüberhinaus den Vorteil, dass sie seit ihrer Erstveröffentlichung ohne große Anpassungen den Technologiewandel weitestgehend überstanden hat. War dieses Projekt ein Beispiel für die detaillierte Analyse und Präsentation einer Quelle und in begrenztem Rahmen ihres Kontextes im Netz, erwuchs zugleich der Wunsch komplementär dazu, eine Infrastruktur zu ermöglichen, die den Zugriff auf größere Quellenbestände mit möglichst umfangreicher technischer Unterstützung ermöglicht. Für die Erforschung der Entwicklung des Wissens zur Mechanik in der frühen Neuzeit wurde 2000 im Rahmen des Archimedes-Projektes [235, 217] mit dem Aufbau einer digitalen Bibliothek begonnen. Diese ermöglicht es Wissenschaftlern aus unterschiedlichen Disziplinen unterstützt durch Sprachtechnologien[7] eine breite Quellengrundlage zu studieren. Hierbei ersetzen Sprachtechnologien nicht die Zusammenarbeit von Historikern, Naturwissenschaftlern und Philologen, sondern sie liefern vielmehr die Grundlage für eine gemeinsame Diskussion und unterstützen so das Erreichen gemeinsamer Erkenntnisse. Sprachtechnologien sind damit ein Instrument der Kooperation [200]. Anforderung an das Projekt war die schnelle Erschließung der Quellen mit dem Ziel der historischen Arbeit, nicht die Detailarbeit der kritischen Edition. Die Leitidee war daher im wesentlichen Funktionen zu implementieren, die den Fragestellungen angemessen waren, d.h. die Hilfe bei der Übersetzung der Texte und Zugriff auf die digitalen Faksimiles zusätzlich zu Transkriptionen. Die grundsätzliche Parallelpublikation von Faksimile und Transkription erlaubte, auf exakte diplomatische Transkription zu verzichten, da bei Problemen der Lesbarkeit oder Streichungen auf das Bild zurückgegriffen werden kann. Verfügbarmachung stand hier in diesem Sinne vor der Genauigkeit, jedoch mit der Möglichkeit, Korrekturen vorzunehmen. Die Verlässlichkeit der jeweiligen Transkription wurde durch ein Ampelsystem symbolisiert und CVS [48] zur Versionierung benutzt. Zur Auszeichnung wurde ein eigener Standard in XML realisiert. Zwar gab es zu diesem Zeitpunkt bereits Diskussionen um die Auszeichnung der Texte im Standard der damals gerade wachsenden *Text Encoding Initiative (TEI)* [232], jedoch wurde damals auf TEI zugunsten eines minimalen Auszeichnungsstandards, der nur die tatsächlich für das Projekt relevanten Auszeichnungen beinhaltet, verzichtet - eine Entscheidung, die man heute eventuell anders treffen würde. Jedoch wäre der Zeitaufwand, eine Teilmenge von TEI zu bestimmen und so genau zu beschreiben, dass eine möglichst eindeutige Auszeichnung sicherstellt gestellt wird, nicht größer gewesen als die Definition und Dokumentation eines eigenen Schemas.

Dazu kam, dass es in der Tat einen zumindest zum damaligen Stand der TEI unterschiedlichen

[6] Eine logische Erweiterung dieses Konzeptes wäre ein Mechanismus, der es dem Rezipienten ermöglicht einen eigenen Weg durch das Manuskript aufzuzeichnen, zu kommentieren und schließlich als eigene Publikation über das Manuskript zu veröffentlichen. In der Tat waren es diese Ansätze die z.B. Peter Damerow bei seinen historischen Arbeiten schon vor HTML und dem WWW mit HyperCard in den 1990er Jahren auf dem Apple Mac realisierte.

[7] Die Texte werden morphologisch analysiert und dann über einen Hyperlink mit relevanten ebenfalls digitalisierten zeitgenössischen Wörterbüchern verbunden.

Ansatz gab, nämlich den Schwerpunkt auf semantische Einheiten eines Textes zu legen, die nicht grundsätzlich mit linguistischen Einheiten übereinstimmen müssen und orthogonal zu diesen stehen können. Ein Problem, dass sich konkret in XML durch den Einsatz von sogenannten Milestones lösen läßt. Zum Beispiel können Seiten- und Zeilenumbrüche als geschlossene leere Tags gekennzeichnet werden, anstatt den Inhalt der Seite in einem entsprechenden Tag einzuschließen.[8] Analog wäre die Umsetzung durch Standoff-Markup möglich gewesen und wurde auch diskutiert, erschien aber zu diesem Zeitpunkt sowohl für die Eingabe als auch für die Webdarstellung zu aufwendig. Der Standard, der letztendlich gewählt wurde, ist jedoch mit dem heutigen TEI-P5 Standard [231] kompatibel und in der Tat werden die Daten schrittweise konvertiert. Schon früh gab es eine enge Kooperation mit dem Perseus-Projekt [184] zur Implementation und Erweiterung der im Projekt genutzten Sprachtechnologien.

2003 führten die Erfahrungen, die mit dem Archimedes-Projekt gemacht wurden, zum Projekt *European Cultural Heritage Online (ECHO)*, einem EU geförderten Projekt, mit dem Ziel eines beispielhaften Aufbaus einer digitalen Publikationsplattform und Arbeitsumgebung zur Bereitstellung des kulturellen Erbes in der digitalen Welt. Von Anfang an war dort einerseits die gemeinsame Entwicklung von Hilfsmitteln zur Analyse der Daten und der Präsentationsumgebung für Quellen, als auch die konsequente Idee des Open Access zu den Quellen konstitutiv. Im Zentrum stand die Idee einer Agora, die Anbieter von Quellen, in diesem Falle Museen, Archive und Bibliotheken, mit den Nutzern der Quellen in Kontakt bringen sollte. Zum Zeitpunkt des Aufbaus von ECHO war jedoch zunächst die Quellenbereitstellung zentral. Da die kritische Masse von Quellen, die für das digitale Arbeiten notwendig ist, damals noch lange nicht erreicht war, wurden hier zunächst Workflows und Tools entwickelt, die für den Aufbau einer digitalen Bibliothek und die Betrachtung von Quellen – im wesentlichen digitale Faksimiles – notwendig waren [69].

Parallel jedoch erstanden sehr spezifische Arbeitsumgebungen, beispielhaft Arboreal, ein Agent zum Browsen durch Textbestände [11], dem Vergleich von Dokumenten sowie zur sprachübergreifenden Annotation von technischen Ausdrücken. Bereits hier stand die Vision im Vordergrund, ein internet-basiertes interaktives Tool zu entwickeln, das die Arbeit mit Texten im Netz, die Kombination unterschiedlicher Ressourcen und die Anreicherung dieser Texte mit interpretativen Annotationen ermöglicht. Auch hier sollte nun technisch Mögliches zur Beantwortung von Fragestellungen aus der Wissenschaftsgeschichte genutzt werden, die vorher so nicht angegangen werden konnten. Insbesondere ermöglichte die Auszeichnung von sogenannten technischen Termini – also Begriffen in Texten, die mit Konzepten assoziiert werden können – die Untersuchung der Ausprägung einer festen Terminologie in wissenschaftshistorischen Kontexten [215].

Neben dem Aufbau eigener Datenbestände steht die Wissenschaftsgeschichte – beispielhaft für andere Disziplinen – vor der Herausforderung, die Datenmengen, die in den modernen wissenschaftlichen Disziplinen im 20. und 21. Jahrhundert produziert wurden und werden, als Quelle zu erschliessen. Dieses macht neue Qualifikationsprofile notwendig und die Entwicklung von Standards zum Austausch von Daten. Umgekehrt stellt sich für die Naturwissenschaften die Frage, welche Daten, die in

[8] Konkret wird <l/> bzw.
 genutzt, um einen Umbruch zu kennzeichnen, anstelle von Konstruktionen wie <page><line>xxxx</line>…<line>xxx</line></page>.

den laufenden Großprojekten, wie etwa in der Umweltforschung, Astronomie und Astrophysik und den Lebenswissenschaften erzeugt werden, in welcher Form archiviert werden sollen, so dass sie für spätere Nachnutzungen weiterhin zur Verfügung stehen. Dieser Punkt wird in dieser Arbeit nur am Rande betrachtet werden, spielt jedoch für Standardisierungsüberlegungen von Schnittstellen und Formaten eine zentrale Rolle.

2.6 Datenmodellierung: Ergebnis geisteswissenschaftlicher Forschung

Im Vergleich zu den Datenbanken aus den Lebenswissenschaften oder der physikalischen Forschung ist häufig nicht die Datenmenge, sondern die Struktur der Daten, insbesondere ihre Fluidität, die Herausforderung an die technische Realisierung. Die Charakterisierung relevanter Daten und ihre Strukturierung, insbesondere die Festlegung, welche Verknüpfungen zwischen ihnen abgebildet werden sollen und wie diese beschrieben werden können, sind ein wesentlicher Teil der Arbeit mit den Daten. Während der Akkumulation der Daten verändert sich die Fragestellung an diese Daten und das hinter der Eingabe stehende Modell ständig.[9] Die Forschungsfrage nach der Strukturierung menschlichen Wissens findet hier ihren direkten Ausdruck in der ständigen Anpassung der Modellierung des Wissens über historische Kontexte in einer Datenbank. In diesem Sinne ist das Modell der Daten Endpunkt und nicht Anfangspunkt der Arbeit mit Datenbanken und -strukturen in den Geisteswissenschaften. Abhängig von Fragestellung und theoretischem Hintergrund können sich Modelle widersprechen oder ergänzen. Es stellt sich also die Frage, wieviel Einheitlichkeit und Standardisierung möglich und sinnvoll ist, wo dieses nützlich ist oder wo dieses im Gegenteil hinderlich ist.

Gegeben die Rolle, die Daten nun auch in den Geisteswissenschaften spielen und die mit ihrer Sammlung, Sichtung und Strukturierung verbundenen intellektuellen Leistungen, ist die Publikation von Daten als wissenschaftliche Leistung immer noch wenig anerkannt. Dieses hat ohne Zweifel Gründe, die tief in der wissenschaftlichen Praxis verankert sind. Jedoch sind es auch technische Hindernisse, die nicht ohne weiteres vernachlässigt werden können. Datenpublikation wird nur dann einen Status als wissenschaftliche Leistung vergleichbar mit Monographien und Aufsätzen gewinnen können, wenn die langfristige Verfügbarkeit gesichert ist. Langfristig in den Zeitskalen der Historiker hat hier eine deutlich andere Bedeutung als in den Naturwissenschaften und erstreckt sich eher über Jahrhunderte als über Jahrzehnte. Daher müssen Verfahren und Standards entwickelt werden, die weitestgehend unabhängig von der technischen Entwicklung sind, – mehr noch: Im Idealfall sollten die wesentlichen Ergebnisse ohne großen technischen Aufwand rekonstruierbar sein. Dabei ist weniger der eigentliche Inhalt der Datenbanken das zentrale Problem. Im Vordergrund steht die grundsätzliche Frage, wie Strukturen und Relationen innerhalb von Datenbanken möglichst nachvollziehbar und einheitlich abgebildet werden können.

Blickt man zurück auf die Transmissionsgeschichte des Wissens, ist der sicherste Weg der Erhaltung und Entwicklung einer Wissensbasis das immerwährende Aufgreifen durch Kopieren und Nach-

[9] Mit diesem Problem beschäftigen sich ausführlich die Arbeiten von Richard Lenz zu „Evolutionären Informationssystemen“. Siehe dazu seine Habilitationsschrift [142], sowie die bisher nicht veröffentlichten Folien seiner Vorlesung an der FAU zum gleichen Thema im WS 2016/2017, für deren Überlassung ich mich hiermit herzlich bei Richard Lenz bedanke.

nutzung in unterschiedlichen Kontexten. Übertragen auf die Datenbank als Wissensbasis stellt sich daher die Herausforderung, Datenbanken so zu strukturieren und zu formalisieren, dass wesentliche Bestandteile ausserhalb des disziplinären Kontextes nachgenutzt werden können und mit Beständen aus anderen Kontexten in Beziehung gesetzt werden können.

Die von der TEI aufgestellten Prinzipien und Standards sind hierbei für textuelle Quellen eine Hilfe. Eine tatsächliche Wiederverwendung von Daten setzt neben ihrer einfachen Verfügbarkeit voraus, dass auch die den Daten zugrundeliegenden Modelle in einer nachnutzbaren Weise zur Verfügung gestellt werden. Hier sind RDF-basierte Ansätze, die sich Ontologien zu Nutze machen, um Daten einerseits allgemein, andererseits jedoch auch disziplinenspezifisch beschreiben zu können, vielversprechend. Das *Conceptual Reference Model*(*CRM*) des *International Committee for Documentation* (*CIDOC*) des *International Council of Museums* (*ICOM*) ist in diesem Kontext ein vielversprechender Ansatz. Darauf werden wir in Kapitel 4.4 zu sprechen kommen.

Sicherlich sind die genannten Herausforderungen nicht grundsätzlich neu und die Möglichkeit der Übertragung von Methoden des Ontology-Engineering, die sich in den letzten zwei Jahrzehnten herausgebildet haben, nun auch auf die geisteswissenschaftliche Methodik deutet auf eine enge Beziehung des informationstheoretischen Wissensmanagements und der Wissensgeschichte als Disziplin in den Geisteswissenschaften. Die Prinzipien, die Thomas Gruber 1995 zu Beginn seines Artikel „Toward principles for the design of ontologies used for knowledge sharing“ zusammenfasst und die als Regeln für die Bildung von Ontologien für ein Wissensmanagement bestimmt waren – 1) *clarity*, 2) *coherence* und 3) *extendibility*[10] – beschreiben letztendlich sehr genau auch die Anforderungen, die an geisteswissenschaftliche Forschung gestellt werden. Die Herausforderung ist die Brücke zwischen der formalen Sprache der Informatik und der Sprache der Geisteswissenschaften zu schlagen. Im Kapitel 5 zur Datenmodellierung komme ich darauf zurück.

[10] Siehe [105, S.909-910].

Kapitel 3

Wissenschaftliche Fragestellung und Anforderungen

3.1 Herausforderungen für die Datenmodellierung

Digitale Daten, verbunden mit Methoden ihrer Auswertung, sind eine neuartige Quellenform für die geisteswissenschaftliche Forschung. Damit unterliegen sie den gleichen Qualitätskriterien wie andere Quellen für die historische Forschung auch. Insbesondere gilt nach wie vor, dass etablierte Methoden der Quellenkritik auch hier angewendet werden müssen. Auch für die Ergebnisse, die auf Grundlage digitaler Forschungsmethoden gewonnen werden, gilt, dass diese mit Ergebnissen aus unterschiedlichen Quellen abgeglichen und gegengeprüft werden müssen. Insbesondere ersetzt die digitale Analyse von Massendaten nicht die Einzelstudie, sondern steht komplementär zu ihr. Der Einsatz von digitalen Analysemethoden kann jedoch insbesondere dazu beitragen, die Relevanz des Einzelfalles innerhalb der Grundgesamtheit besser einzuschätzen.

Zugleich können digitale Methoden ein heuristisches Instrument innerhalb der historischen Forschung sein. Hypothesen können mit digitalen Methoden generiert und auf Plausibilität geprüft werden; sie können dazu beitragen, neue Hypothesen zu generieren, ersetzen aber nicht die Interpretation der Ergebnisse durch die Forscher.

Aus geisteswissenschaftlichen Projekten heraus wiederum werden Datenbanken der wissenschaftlichen Gemeinschaft und der breiteren Öffentlichkeit zu Verfügung gestellt und entwickeln sich daher immer mehr zu einer eigenständigen Form der wissenschaftlichen Publikation. Ein weiterer Aspekt, der zunehmend an Popularität gewinnt, ist verbunden mit der Frage, wie der wissenschaftliche Prozess selbst durch Datenmodelle festgehalten werden kann – einerseits bei der Rekonstruktion historischer Entwicklungen[1], andererseits für die Dokumentation aktueller Forschung in Forschungseinrichtungen und -verbünden.[2]

Das volle Potential von Daten als Publikationsform ist hierbei längst nicht ausgeschöpft. Neben den technischen Problemen bei der Publikation von Daten sind es konzeptionelle Probleme, die gelöst

[1] Frühe Arbeiten dazu finden sich bei [99].

[2] Zum Beispiel in der Erweiterung CRM_{inf} von CRM [117, 249] oder in den Arbeiten im Rahmen des Teilprojektes „Architekturen des Wissens“ innerhalb des Exzellenzclusters „Bild, Wissen, Gestaltung“ in Berlin [12].

werden müssen. Die Heterogenität der Fragestellungen und Methoden innerhalb der Geisteswissenschaften spiegelt sich in einer fehlenden Vereinheitlichung der Daten und Datenmodelle wider. Die Nachnutzung von Daten und die Verknüpfung von unterschiedlichen Datenbanken werden dadurch erheblich erschwert. Darüber hinaus sind weitere grundlegende Voraussetzungen für die Veröffentlichung der Daten als wissenschaftliche Publikation bisher nicht gegeben. Neben der Frage der langfristigen Verfügbarkeit ist dies insbesondere das Problem der Provenienz von Daten in gemeinsam erstellten Datenbanken. Dabei ist das letztere Problem, auch wenn man einen beginnenden Kulturwandel in den Geisteswissenschaften hin zu Open-Access attestieren kann, nicht nur eine Frage des geistigen Eigentums der Daten, sondern vor allem ein Problem der Kontextualisierung, die für die Interpretation der Daten unerlässlich ist.

Es fehlt an allgemein anerkannten Kriterien für die Qualität von Datenpublikationen als Veröffentlichung. Mit der hier vorgelegten Darstellung der Struktur und Methoden geisteswissenschaftlicher Datenbanken sollen im Folgenden Anregungen für solche Kriterien gegeben werden und zugleich Ansätze zu einer Methodik der Quellenkritik für Datenpublikationen aufgezeigt werden.

3.2 Thesen: Netzwerktheorie und Wissenssysteme

Zunächst benötigen wir ein Grundverständnis für die in geisteswissenschaftlichen Datenbanken kodierten Informationen. Hierbei gehen wir davon aus, dass die Erfassung von Daten in Datenbanken immer auf der Grundlage eines spezifischen Wissenssystems geschieht und dass zugleich mit diesen Daten ein solches Wissenssystem beschrieben wird. Diese Wissenssysteme können von sehr unterschiedlicher Komplexität sein – sie können überlappen, deckungsgleich oder vollständig verschieden sein. Strukturierte Daten sind aber niemals ohne die Annahme eines dahinter liegenden Wissenssystems denkbar. Ein Wissenssystem hat „nicht das individuelle Wissen zum Gegenstand [...], sondern das gesellschaftliche Wissen, d. h., [die Frage nach dem historischen Wandel von Wissenssystemen] bezieht sich auf das Wissen nur insofern, als es von zahlreichen Individuen intersubjektiv geteilt und historisch mit einer gewissen Kontinuität tradiert wird.“[52, S.1] Die Rekonstruktion von Wissenssysteme erfordert daher die Analyse und Beschreibung komplexer eng miteinander verflochtener Teilsysteme.[3]

[3] Ausführlich werden diese Strukturen im folgenden Kapitel insbesondere in 5.13 beschrieben werden.

Weder *modellierungstheoretische* noch *netzwerktheoretische* Beschreibungen alleine reichen aus, um die komplexen Strukturen dieser Wissenssysteme zu beschreiben und quantitativ zu analysieren. Dazu bedarf es einer Zusammenführung der beiden Ansätze. Die folgenden Thesen fassen hierbei die Leitgedanken, die hinter dieser Arbeit stehen, zusammen. Nicht auf jede These wird in der Arbeit ausführlich eingegangen werden, sie stehen jedoch im Hintergrund des Forschungsprogramms, in den diese Arbeit eingebettet ist.

Hauptthese (HT): Die Beschreibung von Wissenssystemen erfordert eine Integration von netzwerktheoretischer und modellierungstheoretischer Beschreibung (NMB). *Netzwerktheoretisch* bedeutet hier das Aufgreifen von Methoden der sozialen Netzwerktheorie und der Theorie ökonomischer Verflechtungen im Sinne von Macht, Prestige und Einfluß sowie der Diffusionstheorie in Netzwerken. Die mathematische Formulierung erfolgt hierbei durch die mathematische Graphentheorie, inklusive der diskreten Topologie, durch die lineare Algebra in Form der Matrizendarstellung von Abhängigkeitsbeziehungen, sowie in begrenztem Maße durch Formalismen der diskreten Differentalgeometrie und Analysis.

Modellierungstheoretisch umfasst ebenfalls zwei Ansätze: die Modellierung der semantischen Struktur eines Netzwerkes mittels Ontologien und ihrer formalen Repräsentation, zum Beispiel durch OWL und RDF, sowie die Entwicklung von analytischen Modellen für historische Dynamiken [243].

T1: Netzwerktheorie ermöglicht eine klare Beschreibung von Einflussfaktoren und Abhängigkeiten in komplexen Systemen. Die netzwerktheoretische Beschreibung erfordert eine klare, eindeutige Definition von Knoten und Kanten eines Netzwerkes. Dieses birgt jedoch die Gefahr einer Vereinfachung von Erklärungsmustern, einer Reduktion auf binäre Beziehung (existiert / existiert nicht) und einer Vermischung von Kausalität und Korrelationen als Erklärungsmuster.

T2: Modellierungstheoretische Beschreibungen erlauben eine semantische Definition der Beziehungen in einem Netzwerk und ihrer inneren Abhängigkeiten.

T3: Die Beschreibung von Verflechtungen zwischen unterschiedlichen Netzwerken kann in Form von Multilevel-Netzwerken erfolgen. Diese bergen jedoch die Gefahr einer Simplifizierung der Beziehungen zwischen den unterschiedlichen Netzwerktypen. *Multilevel-Netzwerke*[4] suggerieren eine topologische Aufspaltung der Netzwerke in eine $\mathrm{N x T(N)}$ Struktur, wobei N ein topologischer Raum der Dimension 1 ist und $\mathrm{T(n)}$ prinzipiell voneinander unabhängige topologische Teilräume sind, die lediglich lose miteinander gekoppelt sind.

WD1: Wissensdynamiken folgen einer komplexen Dynamik, wobei die unterschiedlichen Netzwerktypen eng miteinander verschränkt sind. Die Beschreibung der Entwicklung von Wissen

[4] Wir werden *Multilevel-Netzwerke* ausführlicher in Kapitel 5 besonders in Abschnitt 5.8 behandeln. Für jetzt reicht es, hierunter Netzwerke zu verstehen, die aus einer Menge verschiedenen Typen von Kanten und Knoten bestehen, die in Wechselwirkung zueinander stehen.

und dessen Strukturen mit Hilfe von NMB basiert auf der Analyse von komplexen miteinander verflochtenen Netzwerken. Diese Netzwerke werden wir in Form von *semiotischen*, *semantischen* und *sozialen Netzwerken* einführen. Die Netzwerke können grundsätzlich unterschiedliche topologische Strukturen haben. Diese bedingen jedoch einander und sind eng gekoppelt, sie bilden einen topologischen Raum, der nur in Näherung durch Homomorphismen auf $\mathrm{N}\mathrm{x}\mathrm{T}(\mathrm{N})$ abgebildet werden kann. Damit wirkt jede Netzwerkebene als regulativer Mechanismus der jeweils anderen Ebenen.

WD2: Für die Teilnetzwerke existieren Ordnungssysteme und Regulative. Diese sind eng miteinander verbunden und aufeinander abbildbar. Die Beziehungen der Knoten und Kanten innerhalb der Netzwerke werden durch Ordnungssysteme und Regulative eingeschränkt. Ziel des modellierungstheoretischen Ansatzes ist es, für diese Systeme Beschreibungen für eine computergestützte Auswertung zu finden.

WD3: Spannungen im Wissenssystem entstehen, wenn die Topologien der unterschiedlichen Ebenen der Wissensnetzwerke nicht kompatibel sind. Diese sind die Grundlagen einer Dynamik von Wissenssystemen. Dieses kann auch für Teilsysteme in einer Ebene gelten. Hierunter verstehen wir zum Beispiel, wenn die Vernetzungsgrade sich deutlich unterscheiden, oder die Stellung von Knoten in einem Netzwerk sich deutlich von der Stellung der Knoten auf anderen Ebenen unterscheidet, mit denen diese verbunden sind.

WD4: Beziehungen innerhalb von Netzwerken müssen den einzelnen Akteuren nicht bewusst sein. Knoten und Kanten eines Netzwerkes sind voneinander unabhängige Entitäten und werden erst durch Regulative aneinander gebunden. Beziehungen entstehen aufgrund materieller Rahmenvorgaben und sozialer Konventionen.

3.3 Autorschaft, Datenbanken und Datenpublikation

Bei einer kollektiven Publikation stellt sich immer die Frage der Verantwortung und Urheberschaft für bestimmte Teile. Im Falle von Datenbanken[5] ist dieses kein Spezifikum des digitalen Mediums, dies gilt auch für klassische Karteien in Museen und Bibliotheken, hinter denen Generationen von Mitarbeitern und deren Expertise stehen. Werkanalysen mit dem Ziel, die Provenienz einzelner Teile publizierter Werke zu bestimmen, sind eine ständige Herausforderung der historischen, philologischen und philosophischen Beschäftigung mit Quellen, um den Inhalt eines Werkes im Kontext genauer zu verstehen. In der Datenbank wird jedoch die methodische Ausnahme, das Erstellen eines Dokumentes mit mehreren Autoren, zur Regel. Der Beitrag einzelner Autoren ist hierbei hochgradig fragmentiert. Trotzdem ist es von großem Interesse, den Beitrag einzelner Autoren zum Gesamtwerk der Datenbank nachvollziehen zu können. Die Gründe dafür sind offensichtlich und beginnen im einfachen Fall bei methodisch-systematischen Fehlern, die von einem einzelnen Autor in eine Datenbank hineingebracht werden und damit den gesamten Datenbestand kompromittieren können, und enden bei der

[5] Zur Definition von Datenbanken in unserem Kontext siehe Abschnitt 4.1.

für die Datenbank als Publikationsform wesentlichen Möglichkeit, unterschiedliche Auffassungen zu einzelnen Einträgen zu dokumentieren.

Neben dem Interesse des Nutzers einer Datenbank, den Autor eines Beitrages erkennen zu können, ist umgekehrt für den Autor die Möglichkeit, auf seinen Anteil an einer Publikation verweisen zu können, für die Erhaltung seines geistigen Eigentums unumgänglich, nicht zuletzt aus Gründen der wissenschaftlichen Reputation. Die Sicherung der Autorschaft ist die Voraussetzung für eine frühzeitige Publikation von Daten.

3.4 Ereignisse als zentrales Konzept digitaler Repräsentation

Schauen wir uns den Umgang mit Daten und deren Akkumulation näher an und berücksichtigen Aspekte wie unterschiedliche Auffassungen zu demselben Gegenstand, so lässt sich eine Gemeinsamkeit deutlich ausmachen. Fast immer lässt sich – zugegeben sehr vereinfachend –die Zuordnung einer Eigenschaft zu einem Gegenstand in der Form „*X meint über Y zum Zeitpunkt T das Folgende*“ ausdrücken verbunden mit dem Verweis zu einem Kontext, der die Gründe dieser Zuordnung umfasst [117]. Die Sammlung all dieser Aussagen zu einem Gegenstand ergibt eine Übersicht über den aktuellen Stand des Wissens zu diesem Gegenstand. Mit diesem Modell lassen sich unterschiedliche Auffassungen zu einem Gegenstand und Veränderungen von Zuordnungen unmittelbar beschreiben. Eine ereignisorientierte Beschreibung ergibt sich so in natürlicher Weise. Als *Ereignisse* verstehe ich hierbei Entitäten, die in Zeit und Raum verortet werden können und die von einem oder mehreren Akteuren beeinflusst worden sind oder diese beeinflussen.[6]

Hierbei können konkrete Artefakte, Personen oder Orte, genauso aber auch abstrakte Konzepte der Gegenstand sein. Versieht man diese Aussagen mit Kontexten, in denen sie gefasst wurden, so lassen sich diese Ereignisaussagen gruppieren, negieren und in andere Bezüge setzen. Widersprüchliche Aussagen können festgehalten werden.

Die Erfassung von Eigenschaften eines Gegenstandes in der Form von Zuordnungen bzw. in Form von Aussagen über diesen Gegenstand ermöglicht es, diese in natürlicher Art und Weise auf innere Konsistenz zu prüfen und gegebenenfalls zu korrigieren.

Aussagen, die irrtümlich *einem* Gegenstand zugeschrieben wurden, können aufgespalten werden und dann auf unterschiedliche Objekte verteilt werden. Zu Unrecht als unterschiedlich angesehene Entitäten können zusammengefasst werden, ohne dass die Struktur der Aussagen grundlegend geändert werden muss. Insbesondere kann die Zusammenführung von Entitäten durch Aussagen wie: wie „*A ist identisch zu B*“, nachvollziehbar und umkehrbar festgehalten werden.

Die Beschreibung von Eigenschaften einer Entität in Form von Aussagen legt eine Formalisierung durch RDF mit seinen Erweiterungen RDFS und OWL nahe.[7] Das Dokumentieren von Veränderungen oder differierenden Meinungen in der Beschreibung von Eigenschaften von Objekten in einer tabellenorientierten Datenbank erfordert hingegen einen höheren Aufwand. Entweder muss für jede

[6] Das heisst nicht, dass all diese Informationen bekannt sein müssen.

[7] Auf RDF, RDFS und OWL gehen wir in 4.2 näher ein.

Veränderung eine Teilkopie des Datensatzes mit entsprechenden Metadaten[8] über die Veränderung des Datensatzes angelegt werden oder es werden entsprechend aufwendige Kreuztabellen für einzelne Felder erstellt. Letztendlich stellen diese Lösungen nichts anderes als eine Implementation von RDF in tabellenorientierten Systemen dar.

Ein weiteres starkes Argument für Ereignisse als zentrales Konzept der Strukturierung von Datenbanken für die geisteswissenschaftliche Forschung, insbesondere für die historische Forschung, ist der betrachtete Gegenstand selbst. Ereignisse sind der grundlegende Baustein für jede Analyse historischer Prozesse. Die ereignisbasierte Modellierung vereinfacht daher den Transformationsprozess von der wissenschaftlichen Fragestellung in ein Datenmodell erheblich. Diese Form der Modellierung wird hiermit zu einer gemeinsamen Sprache von Informatik und Geisteswissenschaft und vereinfacht so den Austausch zwischen diesen Disziplinen.[9]

3.5 Aussagen und Aussagen über Aussagen

In Abschnitt 2.3.1 hatten wir den Forschungszyklus als Motivation für die folgenden Modellierungsansätze vorgestellt. Dieser findet sich hier wieder, indem wir die Iterationen des Zyklus als Systeme von Aussagen interpretieren, die sich in jedem Zyklus weiterentwickeln und sich dann als Aussagen über Aussagen in unserem Datenmodell wiederfinden. In jedem Schritt entstehen hierbei Mengen von Daten, die in einem Kontext zusammengefasst werden können. Diese kommentieren die in den Schritten vorher angelegten Aussagen oder fügen neue Aussagen hinzu. Für jede dieser Iterationen gelten grundsätzlich andere Annahmen über die Konsistenz und innere Logik von Daten, die von der Forschungsfrage und Methodologie abhängen. Dies hat konkrete Auswirkungen auf die Modellierung dieser Schritte in einem Datenmodell. In 3.1 werden wir dieses weiter ausführen. Selbstverständlich lassen sich nicht alle Aussagesysteme als Ergebnis geisteswissenschaftlicher Reflexion als formal logische Systeme modellieren. Diese Herangehensweise kann und will nur einen sehr spezifischen Teil der Auswertung und Reflexion wiedergeben. In der Praxis besteht die Bearbeitung und Analyse von Texten aus einem komplexen Geflecht der Quellenbearbeitung, zu dem die freie Assoziation genauso gehört wie die penible Rekonstruktion. Dies spiegelt sich in der Diskussion um die Frage, was Annotation von Dokumenten bedeutet, genauso wider wie in der Frage der Abgrenzung zwischen Daten und Metadaten. Auf letzteres werden wir in 4.1.1 etwas genauer eingehen.

[8] Zu Daten und Metadaten siehe 4.1.1.

[9] Zur Umsetzung dieses Konzeptes siehe 3.4 und 4.4.

Kapitel 4

Modellierung und Datenbanken

Auch aus meiner Sicht war es lange ein Fehler, ein auf den ersten Blick zunächst neues Konzept innerhalb der Informatik unter den in den Geisteswissenschaften – insbesondere der Philosophie und historischen Forschung – etablierten Begriff der *Ontologie* zu fassen. In der Tat hat die Kritik von Seiten der Geisteswissenschaften im Hinblick auf die Verwendung dieses Begriffes nicht lange auf sich warten lassen.[1] Die langjährige Beschäftigung mit der Frage der Rolle von Datenbanken und computergestützter geisteswissenschaftlicher Forschung haben mich mittlerweile jedoch überzeugt, dass diese Begriffswahl letztendlich eine sinnvolle war und – eine Offenheit von beiden Seiten vorausgesetzt – die Chancen für den Dialog zwischen den Disziplinen verbessert. Das Konzept Ontologie ermöglicht es oder vielmehr zwingt dazu, zwei sich widersprechende Positionen bzw. Ziele auf einer grundsätzlichen Ebene zu hinterfragen: Auf der einen Seite besteht die Hoffnung, menschliches Wissen in formale Strukturen zu bringen, mit Hilfe von Computern bewertbar und dessen Entwicklung berechenbar zu machen. Auf der anderen steht die Überzeugung von der unvorhersagbaren Kreativität und Unschärfe von Erkenntnisprozessen, die eine Vorhersagbarkeit der Wissensentwicklung mit Hilfe von Berechnungen durch Computer unmöglich machen. Die frühen Diskussionen um Ontologiebildung für das Wissensmanagement in der ersten Hälfte der 90er Jahre des letzten Jahrhunderts, wie etwa die schon in der Einleitung[2] erwähnten Arbeiten von Thomas Gruber [105], zeigen, dass dieses Problem zwar von der Seite der Informationstheorie erkannt wurde, jedoch in die geisteswissenschaftliche Diskussion so gut wie keinen Eingang gefunden hat. Erst die unmittelbare Notwendigkeit, auch in den Geisteswissenschaften große Mengen von Daten zu verarbeiten, zwingt diese nun in den Dialog.

Doch es gilt auch, die Unterschiede der beiden Ansätze deutlich herauszuarbeiten und zu vermitteln. Während die Ontologie in der Philosophie den Anspruch hat, die Strukturen zu erfassen, die für das Sein konstitutiv sind, verfolgt das Ontologiekonzept Grubers einen reduktionistischen Ansatz. Die Ontologie als Struktur ist als formales System gesetzt. Es besteht in diesem Sinne kein Vollständigkeitsanspruch. Ziel ist vielmehr die widerspruchsfreie Beschreibung der Dinge, die im Zielbereich der Ontologie liegen, und damit die Möglichkeit einer Klassifikation von Entitäten auf der Grundlage präzise definierter Eigenschaften. Die Geisteswissenschaften, und insbesondere diejenigen Disziplinen,

[1] Für eine kurze Einführung in die Begriffswelt und Diskussion siehe [106] und die dort zitierte Literatur.

[2] Siehe Abschnitt 2.6.

die sich mit der Geschichte des Wissens und seiner Strukturen befassen, haben hier eine besondere Funktion als Mittler zwischen den beiden Ansätzen. Die Rekonstruktion der eingangs geschilderten semiotischen Netzwerke in ihrem jeweiligen historischen Kontext ist in diesem Sinne der Versuch, die formale bzw. formalisierte Beschreibung von Wissen, die implizit oder explizit hinter den *semiotischen Netzwerken* steht, mit den ontologischen Konzepten, die die Strukturen des semantischen Netzwerkes bestimmen, in Beziehung zu setzen und ihre Koevolution zu erklären. Diese semantischen Ontologien zielen zumindest dem Anspruch nach darauf ab, menschliches Wissen vollständig zu strukturieren, und stehen damit in der Tradition des philosophischen Ontologiebegriffes. Die im sozialen Netz repräsentierten sozialen Akteure sind hierbei die treibende Kraft für die miteinander korrespondierenden Veränderungen der beiden ontologischen Systeme. Den Regeln der Hermeneutik kommt hierbei eine zentrale Rolle zu. Sie setzen den Rahmen für die Verfahren der Rezeption von Wissenssystemen über die Interpretation der Manifestation von Wissen in Zeichensystemen – in unserem Kontext den *semiotischen Netzwerken* –, zugleich aber auch für die Neuorganisation des Netzwerkes durch den Rezipienten, der als Ergebnis seiner Reflexion vor dem Hintergrund seines Wissens – und damit des inkorporierten *semantischen Netzwerkes* – eine Neuordnung des Zeichensystems vornimmt. Der klassische hermeneutische Zirkel findet hier einen expliziten Ausdruck.

Von der praktischen Seite ist die Aufgabe des *Tabellenparadigmas* ein entscheidender Schritt in die Richtung, Datenstrukturen zu entwickeln, die disziplinenübergreifend auf Akzeptanz stoßen können. Die Erfassung von Daten in Tabellen erfordert grundsätzlich eine strikte Einordnung von Objekten als Instanzen einer Klasse, bevor eine Erfassung, Zuordnung und Beschreibung in einer Tabelle erfolgen kann. Natürlich ist es prinzipiell möglich, Eigenschaften zu verändern, Daten von einer Tabelle in eine andere zu verlagern und das Datenschema auf der Grundlage neuer Anforderungen zu verändern. Dies erfordert in der Regel jedoch einen tiefen Eingriff in die Daten sowie einen doch häufig erheblichen Programmieraufwand und ist mit Risiken des Informationsverlustes verbunden.[3] Datenerfassung in der geisteswissenschaftlichen[4] Forschung verfolgt jedoch häufig genau den im Tabellenparadigma schwer zu realisierenden Weg, Objekte und deren Eigenschaften lediglich vorläufig zu klassifizieren und im Verlauf des Forschungsprozesses Eigenschaftsbeschreibungen und Klassifikationssysteme fortlaufend zu verändern. Darüber hinaus ist eine eindeutige Klassifikation von Objekten oft nicht möglich.

Hinter dem Tabellenparadigma steht grundsätzlich eine feste Verknüpfung von Eigenschaften mit einer Klasse. Ein Objekt mit Eigenschaften, die unterschiedlichen Klassen zugeordnet werden können, erscheint entweder in verschiedenen, in keiner oder in einer neuen Tabelle mit Feldern, die sich durch das Zusammenlegen mehrerer Klassen ergeben. Hierbei ist unklar, welches am Ende die bessere Lösung ist. Natürlich lässt sich diese Aufgabe mit einem System von Querverweisen, Kreuztabellen und Schlüsseln lösen. Immer entstehen hierbei jedoch mehrfache Instanziierungen des gleichen Objektes, die nur notdürftig zusammengehalten werden. Alle zu einem Objekt erfassten

[3] Der Aufwand Objekte neu zu klassifizieren oder Objekte auf unterschiedliche Tabellen mit unterschiedlichen Schemata zu verteilen, wird in der Praxis häufig gescheut. Das Ergebnis ist in der Regel das Aufblähen der Tabellen um immer weitere Felder, wie sich gut an der Entwicklung von Standards wie MIDAS oder auch Lido und MARK nachvollziehen lässt.

[4] Die vorgenommene Einschränkung auf die Geisteswissenschaften hier ist dem Thema der vorliegenden Arbeit und nicht meiner Überzeugung geschuldet, dass hier auch Wissenschaft stehen könnte.

Daten in einem komplexen System zusammenzuführen, das durch Schlüssel und Fremdschlüssel zusammengehalten wird, ist häufig kompliziert und mit dem Risiko verbunden, Informationen über die Bedeutung von Relationen lediglich im Frontend zu implementieren. Letztendlich sind diese Lösungen systemfremd und nicht mehr als die Einführung eines ontologiebasierten Ansatzes durch die Hintertür. Der Widerstand der Geisteswissenschaftler gegen einen solchen methodischen Zwang ist nachvollziehbar.

Der Ansatz einer Beschreibung von Datenstrukturen mit Hilfe einer Ontologie löst die geisteswissenschaftliche Forschung aus diesem Methodenzwang. Die Gültigkeit von Eigenschaften in Form von *Definitonsbereich* und *Wertebereich*, *Kardinalitäts-* und *Ausschließungsbedingungen* kann zwar eingeschränkt werden und die Eingabe von in diesem Sinne nicht erlaubten Relationen durch ein entsprechend realisiertes Frontend unterbunden werden. Dies ist jedoch nur dann möglich, wenn die Arbeitsreihenfolge erst eine Klassifizierung erzwingt und dann die Eingabe von Eigenschaften ermöglicht. Genau diese Reihenfolge ist jedoch in ontologiebasierten Systemen nicht zwangsläufig. Es können zunächst Eigenschaften eines Objektes beschrieben werden, bevor dem Objekt eine Klasse zugeordnet wird. Es können Instanzen erzeugt werden, die Eigenschaften haben, die entweder einer bestimmten Klasse nicht zugeordnet werden können oder der eigentlichen Klassendefinition widersprechen oder auch solche, die Eigenschaften mehrerer Klassen haben.

Im Falle der Nicht-Entscheidbarkeit im Moment der Dateneingabe, etwa aus Mangel an vorliegenden Informationen oder wegen grundsätzlicher Fehler in der Definition der Gültigkeitsbereiche von Eigenschaften, ist eine Instanz im Zweifel zunächst eine Instanz von mehreren Klassen, sofern diese nicht sich ausschließende Merkmale haben, bzw. einer Oberklasse – im Extremfall lediglich der Wurzelklasse.[5]

Beim Versuch, einem Objekt eine spezifischen Klasse in Form einer *is_a* Relation zuzuordnen, entstehen gegebenenfalls Widersprüche zum vorgegebenen System. Mit diesen Widersprüchen umzugehen, ist im Sinne eines pragmatischen Ansatzes die Aufgabe der gemeinsamen Diskussion über die vorgegebene Ontologie. Eine Frage, die hierbei anzugehen ist, ist, welcher Grad von Widerspruchsfreiheit für ein System als sinnvoll angesehen wird. Hier liegt ein grundlegender Unterschied von ontologie- und tabellenbasiertem Ansatz.

Ein ontologiebasiertes System gibt somit zunächst nur mögliche Eigenschaften einer Instanz vor. Die in einem Wissenssystem hinterlegte Ontologie schränkt die Vergabe von Eigenschaften prinzipiell ein; neue Eigenschaften müssen dem System immer erst in Form der Erweiterung der Ontologie mitgeteilt werden. Dies ist, da die Ontologie selbst Teil der Wissensbasis ist, einfach möglich, bleibt jedoch eine bewusste Entscheidung, und ein willkürliches Ausweiten eines Beschreibungssystems wird erschwert.

Die Offenheit des Systems wirft unmittelbar die Frage auf, wie mit Standardannahmen über Objekte umgegangen werden soll und kann. Bei klassischen tabellenorientierten Eingabesystemen geschieht hier eine Vorgabe häufig durch die Gestaltung der Eingabeoberfläche: Eingaben von bestimmten Werten werden erzwungen oder in der Tabelle selbst wird ein Defaultwert für bestimmte Werte

[5] In der Praxis wird bei klassischen Datenbanken hier häufig zum Kommentarfeld gegriffen und vermerkt, „Eigentlich stimmt das so nicht, außerdem weiß ich noch…“ usw.

vorgeben, sobald eine neue Instanz angelegt wird.

Für ontologiebasierte Systeme ist dies zunächst in der gleichen Weise über das Oberflächendesign möglich, widerspricht aber dem Ansatz, möglichst alle Annahmen über die Struktur der Daten in der Wissensbasis zu kodieren. In der Objekt orientierten Repräsentation der *Functional Requirements for Bibliographic Records* in *FRBRoo*[6] wurden hierfür *Metaeigenschaften* der Form *should have* vorgeschlagen, die aufgerufen werden, sobald eine Instanz einer Klasse zugeordnet werden soll.[7]

Noch stellt sich dieses Problem als ein Desiderat dar. Der Rückgriff auf das aus den Kognitionswissenschaften stammende Konzept der *mentalen Modelle* [124] und *frames* [154] kann zur Lösung an dieser Stelle hilfreich sein. Erfassen wir Klassen einer Ontologie als Instanziierungen eines mentalen Modells, eröffnet sich hier die Möglichkeit einer klaren Beschreibung des Prozesses der schrittweisen Anreicherung von Informationen an ein Objekt. Nicht bekannte Eigenschaften einer Instanz werden zunächst mittels der durch das mentale Modell vorgegebenen Standardannahmen initialisiert und dann entsprechend auf Grund von neuen Erkenntnissen ergänzt bzw. verändert. Die mentalen Modelle ließen sich in diesem Kontext als Metaklassen mit der Sprache der formalen Ontologie ausdrücken. Die auf den Instanzen der Metaklassen der zugelassenen logischen Operationen können durch die durch *frame logic* vorgegebenen Methoden [7] und durch Methoden aus der Theorie *nichtmonotonen Schließens* erweitert werden [92].

Dieser Ansatz ist aus streng formaler Sicht sicher problematisch, da damit der beschreibungslogische Ansatz zunächst aufgegeben wird. Von der operativen Seite ermöglicht dieser Ansatz jedoch ein anwendungsnahes Herangehen an das grundsätzliche Problem der Erfassung und Beschreibung unvollständiger Entitäten. Auch wenn sich diese Arbeit in den praktischen Teilen im wesentlichen auf die Arbeit mit Ontologien und Instanzen beschränkt, ist die Möglichkeit der Ausweitung des Konzeptes auf diese Form von Metamodellierung jedoch eine wesentliche Motivation für den Einsatz von formalen Ontologien. Diese gilt insbesondere im Rahmen der historischen Epistemologie, dient doch das Konzept der mentalen Modelle auch innerhalb dieses Ansatzes der Wissenschaftsgeschichte als ein Erklärungsmodell für die Stabilität von erklärenden Mustern [203, 202].

Welche Ansätze für Ontologien sind jedoch für die historische Forschung tatsächlich sinnvoll? Welche Kriterien können dazu beitragen, Richtlinien für solche Ontologien zu definieren? Gibt es Hoffnung, eine Referenzontologie zu bestimmen, in die sich andere domänenspezifischen Ontologien einbetten können?

Diese Fragen lassen sich nicht abstrakt beantworten, sondern müssen auf Fallbeispiele und insbesondere auf Forschungsfragen bezogen werden, die in der historischen Forschung relevant sind. Im Folgenden soll dargestellt werden, wie sich generelle Anforderungen an die Datenmodellierung aus konkreten Projekten ableiten lassen. Hierbei ist die Überzeugung des Autors, dass sich aus den Fallbeispielen tatsächlich allgemeingültige Prinzipien und ein Ansatz für eine allgemeine Ontologie ableiten lassen. Für diesen pragmatischen Ansatz ist es notwendig, sowohl Forschungsfragen an große Datenmengen als auch prototypische Anwendungen für die Datenmengen näher zu beleuch-

[6] Siehe [18] sowie auch die Abschnitte 4.4 und 9.2.

[7] Zum Beispiel CLP2, CLP43, CLP45, definiert in [248].

ten.

4.1 Daten und Modelle - eine Arbeitsdefinition

In den vorherigen Kapiteln[8] sind die Begriffe *Daten* und *Datenbank* häufiger gefallen, jedoch ohne eine nähere Begriffsbestimmung. Dieses soll hier nachgeholt werden.

4.1.1 Daten, Metadaten und Datenbank - eine Arbeitsdefinition

Unter *Daten* verstehen wir mit Hilfe von Computern verarbeitbare als Bitstream vorliegende Objekte. Dieses können insbesondere digitale Repräsentationen von Bildern, Texten, Videos, Ziffern usw. sein.[9] Wir benutzen hier bewusst einen engeren Datenbegriff, indem wir uns auf elektronische Daten einschränken. Nicht elektronisch vorliegenden Daten fallen im folgenden unter die Kategorie der *physischen Objekte*.

Metadaten sind Daten, die entweder selbst Daten oder ein physisches Objekt beschreiben. Sie enthalten immer eine Referenz auf diese Daten, auf ein oder mehrere physische Objekte oder auf Sammlungen von Daten oder physischen Objekten. Diese Referenz soll im Allgemeinen einen eindeutigen Rückschluss auf das beschriebene Objekt oder die beschriebenen Objekte erlauben. Metadaten sind damit immer Daten und Metadaten können Metadaten von Metadaten sein. Auch Annotationen oder im Extremfall auch vollständig (digital) vorliegende Texte können in diesem Sinne Metadaten darstellen, solange sie explizit auf anderen Objekte verweisen.

Eine *Datenbank* ist eine Sammlung von Daten, die einen gemeinsamen oder logischen Bezug haben [107, S.615]. Datenbanken sind in diesem Kontext klassische SQL-Datenbanken, XML-Dokumente, Datenbanken, die in einem Triplestore vorliegen können, aber auch digitale Text- oder Bildersammlungen.

4.1.2 Datenmodellierung und Typen von Datenbanken

Unter einem *Datenmodell* fassen wir einen Satz von Metadaten auf, der die formal-logische Struktur einer Datenbank beschreibt. Im praktischen Einsatz ist dies im Falle einer klassischen SQL-Datenbank das Schema, mit dem die Tabellen und ihre internen Verknüpfungen erzeugt werden können, verbunden mit einer inhaltlichen Beschreibung dieser Tabellen, ihrer Felder und Verknüpfungen. Dies kann durch Texte, Graphiken, *Unified Modeling Language* (*UML*, [260]) und ähnliche Formate geschehen. Im Falle von Daten in XML ist dies das kommentierte Schema, das den XML-Dokumenten zugrunde liegt, üblicherweise formal als *Document Type Definition* (*DTD*), *XML-Schema* [254] oder *relax-ng* [199] vorliegend. Im Falle von Daten, die in einem Triplestore gespeichert werden, gehen wir von einer Beschreibung der Datenstruktur in RDF oder OWL aus.[10]

[8] Siehe insbesondere 1.1.1ff, 2.3.1ff und 3.1ff.

[9] Siehe dazu auch die Lehrbuchliteratur, z. B.[107, S.92].

[10] Siehe 4.2.

In den 1970er Jahren wurde von Codd [107, S.619] das Konzept *relationaler Datenbanken* eingeführt. Für die formale Definition sei auf Lehrbuchliteratur verwiesen.[11] Eine Erweiterung dieses Modells erfolgte später durch das Konzept der objektrelationalen Datenbanken [244].

Als Abfragesprache für relationale Datenbanken hat sich *SQL* verbunden mit einzelnen implementationsabhängigen Dialekten und Erweiterungen etabliert. Auch darauf soll hier nicht weiter eingegangen werden.[12]

Neben den SQL-Datenbanken haben sich in den letzten Jahren Datenbanksysteme durchgesetzt, die nicht primär auf SQL als Abfragesprache setzen [149]. Für unseren Anwendungskontext sei neben den Triplestores, die im Folgenden eingeführt werden werden, nur der sehr spezifische Typ von textorientierten und indexbasierten Datenbanken erwähnt, der es ermöglicht, schnell innerhalb großer Textbestände zu suchen. Dominierend sind hier *Apache-Solr* [9] und *Elastic-search* [173]. Beide Datenbanken haben ähnliche Spezifikationen und zeigen vergleichbare Ergebnisse bei Performance-Tests [10]. Ihnen gemeinsam ist, dass sie auf Grundlage unterschiedlicher konfigurierbarer Algorithmen Indizes von Dokumenten erstellen, die dann nach verschiedenen Kriterien abgefragt werden können. Dokumente sind hierbei in der Regel Texte, die tokenisiert und mit Metadaten versehen werden.[13]

4.1.3 Triplestores und Graphdatenbanken

Wir konzentrieren uns hier auf solche Systeme, deren primäre Strukturen Relationen sind. Formal besteht jede Einheit in diesen Datenbanken aus einem Tripel (s,p,o),[14] wobei alle drei Werte des Tripels eindeutige Referenzen darstellen. Diese einfache Form der Daten wurde schließlich noch um einen weiteren Wert (c) erweitert, der es ermöglicht, jedem Tripel einen Kontext zuzuweisen. De facto hat sich damit bei den meisten herkömmlichen Systemen eine Quadrupelstruktur (c,s,p,o) ergeben. Die Behandlung des Kontextes ist hierbei in den unterschiedlichen Implementationen nicht immer einheitlich. Wir werden dieses Strukturen in den folgenden Abschnitten noch ausführlicher behandeln. Neben den Triplestores liegt mit der Graphdatenbank Neo4j [159] noch eine andere Datenbankform vor, die einerseits eine Tripelstruktur erlaubt, andererseits können jedem der Einträge des Tripels Attribute zugewiesen werden; damit werden Eigenschaften tabellenorientierter und graphenorientierter Darstellung vereint. Für die praktische Arbeit haben wir uns für einen Triplestore als Datenbanksystem entschieden. Neo4j und andere Nicht-SQL-Datenbanken kommen hierbei jedoch als sekundäre Systeme vor allem aus Gründen der Performanz ebenfalls zum Einsatz. Auf die konkrete Umsetzung gehen wir in 6.6 ein.

[11] Neben dem schon zitierten [107], z.B. [218] für eine sehr praxisorientierte und zugleich theoretische Einführung.

[12] Auch dazu sei auf die umfangreiche Literatur verwiesen, z.B. [81].

[13] Siehe auch 6.3.

[14] Diese werden in der Regel als Subjekt,Prädikat und Objekt der Prädikatenlogik aufgefasst, daher auch die hier gewählten Abkürzungen: s,p,o.

4.2 Graphen und RDF

Die Grundstruktur für alle unsere weiteren Betrachtungen bildet die Darstellung unserer Datenstrukturen als Graph. Graphen beschreiben wir formal in dem aus den Bemühungen um das *semantische Web* hervorgegangenen *Resource Description Framework* RDF. Für die formale Definition und die möglichen Implementationen sei auf die Dokumentation des WWW-Konsortiums [196] verwiesen. Für unsere Betrachtungen ist hierbei relevant, dass RDF einen Rahmen vorgibt, in dem Relationen zwischen Objekten beschrieben werden können.

4.2.1 Resource Description Framework (RDF)

Die Strukturen und Einsatzmöglichkeiten des *Resource Description Framework* (RDF) werden in der Lehrbuchliteratur umfänglich diskutiert. Daher fasst dieser Abschnitt nur die in unserem Kontext relevanten Aspekte zusammen. RDF ist ein offizieller Standard des W3C und daher in der dort üblichen Weise ausführlich dokumentiert [196]. Die zentrale Funktion von RDF ist auf der Einstiegsseite zu RDF beim W3C zusammengefasst:

> RDF extends the linking structure of the Web to use URIs to name the relationship between things as well as the two ends of the link (this is usually referred to as a “triple”). Using this simple model, it allows structured and semi-structured data to be mixed, exposed, and shared across different applications.

Im Wesentlichen gibt RDF einen Standard vor, wie eindeutig identifizierbare Objekte miteinander in Beziehung gesetzt werden können. Wesentlich hierbei ist, dass die Verbindung zwischen Ressourcen jeweils durch ein Tripel der Form *(Subjekt,Prädikat,Objekt)* beschrieben wird. Alle drei Bestandteile sind hierbei zunächst lediglich eindeutige Identifikationsmerkmale – in der Regel *Unified Resource Identifiers (URI)*. Diese sollten mit Ressourcen verbunden sein. Subjekt und Objekt beziehen sich in der Regel direkt auf Webseiten oder Daten. URIs sind in diesem Falle URLs, die aufgelöst werden können und Zugriff auf diese Ressourcen geben. Gleiches gilt für das Prädikat, hier sollte jedoch auf ein Vokabular verwiesen werden, das die Prädikate näher bestimmt. RDF selbst gibt ein minimales Vokabular vor, das eine wenige, aber essentielle Aussagen über Dokumente im Netz ermöglicht. Der wichtigste Ausdruck hierbei ist das Prädikat **rdf:type** oder in der Kurzform **a** bezeichnet, das einer Ressource eine Klasse zuweist. Außerdem stellt RDF Ausdrücke zur Verfügung, die im Sinne einer Reifikation Aussagen über Aussagen ermöglichen. Dieses sind **rdf:Statement**,**rdf:subject**, **rdf:object** und **rdf:predicate**.

Eine Menge solcher Tripel wird als *Graph* bezeichnet. RDF wird in verschiedenen Implementationen realisiert. Wir verwenden hier in der Regel die Darstellung von RDF, wie sie in RDF 1.1 TriG [194] festgelegt ist. Dieses Schreibweise hat den Vorteil, unmittelbar in die Abfragesprache für Triplestores SPARQL überführt werden zu können.

Die erste Erweiterung des Vokabulares erfolgte durch RDF-Schema; auch dieses ist eine offizielle Empfehlung des W3C [195]. Dieses Vokabular stellt Typdefinitionen wie **rdfs:Class** und **rdfs:Property** zur Verfügung. Die wesentlichste Ergänzung ist die Möglichkeit, durch **rdfs:range** oder **rdfs:domain**

die Geltungsbereiche von Prädikaten, die Eigenschaften (**rdfs:Property**) von Klassen beschreiben, genauer zu fassen.

4.2.2 OWL

Ziel von RDF war die Beschreibung einfacher semantischer Beziehungen zwischen Ressourcen. Ressourcen waren zunächst im Wesentlichen digitale Repräsentationen von komplexen Dokumenten. Die wachsenden Menge von Ressourcen, die Daten in kleinere Sinnabschnitte aufgelöst zur Verfügung stellen, wie etwa Normdaten oder auch Textfragmente und Tabellen aus Datenbanken, machten eine wesentlich detailliertere Beschreibung des Verhältnisses dieser Daten untereinander notwendig, insbesondere wenn eine maschinelle Auswertung der Daten bzw. des Netzwerks der Daten etwa mit Hilfe von Reasonern[15] ermöglicht werden sollte. Mit der Web Ontology Language (OWL)[16] wurde dieses umgesetzt. OWL ermöglicht eine Klassifizierung von Daten in einer Form, so dass logische Operationen und insbesondere Inferenz möglich werden. OWL selbst steht in verschiedenen Formen zur Verfügung. OWL-DL hat explizit das Ziel, nur Aussagen, die beschreibungslogisch valide sind, zuzulassen.[17] Damit sind Aussagen, die in OWL beschrieben sind, immer entscheidbar. Eine komplette Einführung in OWL findet sich mittlerweile in der Standardliteratur [110], die Breite der Anwendungsmöglichkeiten wird in [225] dargelegt.

4.2.3 *Named Graphs*

Ursprünglich auf RDF aufbauenden Ansätzen und beschreibungslogisch motivierten Erweiterungen von RDF zur OWL fehlt die Möglichkeit, Aussagen über Aussagen unmittelbar zu formulieren, ohne den beschreibungslogischen Rahmen zu verlassen. Zwar ermöglicht Reifikation formal die Beschreibung von Konstrukten, die Aussagen über Aussagen nahekommen, die Trennung der unterschiedlichen semantischen Ebenen von Aussagen und Aussagen über Aussagen wird damit jedoch aufgehoben und es können Aussagen formuliert werden, die nicht mehr entscheidbar sind. Betrachten wir dieses Konstrukt unter einem geisteswissenschaftlichen Blickwinkel, so unterscheiden sich die beiden Systeme – Aussagen und Metaaussagen – prinzipiell.[18] Wir erwarten in beiden Systemen grundsätzlich unterschiedliche Regelsysteme. Durch die Reifikation wird diese prinzipielle Trennung aufgegeben. Vereinfacht ausgedrückt, wird Faktenbasis und ihre Interpretation auf eine semantische Ebene gesetzt. Beide Systeme besitzen jedoch in der Regel unterschiedliche Prinzipien, nach denen Widerspruchsfreiheit und Konsistenz überprüft werden müssen. Durch Reifikation wird die Überprüfung der Widerspruchsfreiheit, gemessen an unterschiedlichen Regelsätzen, mit Hilfe von *Reasonern* erschwert. Geht man stattdessen zu *named graphs* und damit faktisch zu Quadrupeln über, bleiben alle diese Möglichkeiten erhalten. In der Quadrupelformulierung fassen wir alle Aussagen, die zu einem System gehören, durch einen Kontext zusammen. Dieses kann entweder mittels *named graph* geschehen, d.h. wir geben allen zusammenhängenden Aussagen den gleichen Kontext, identifiziert

[15] Wie etwa Pellet [183], HermiT [109] oder FaCT++ [178].
[16] Zur Einführung in den vollen Sprachschatz siehe [8].
[17] Zur Einbettung von OWL in den beschreibungslogischen Kontext siehe [93, Kap 4.3].
[18] Siehe Abschnitt 3.5.

durch eine URI, oder wir geben jedem Tripel eine eigene URI und fassen anschließend eine Menge dieser URIs wiederum zusammen. Ein Kontext ist dann durch diese Menge gegeben. Auch wenn diese beiden Ansätze nicht vollständig äquivalent sind – Im letzteren Falle sind alle Aussagen erst einmal kontextfrei und werden erst durch Anbindung kontextualisiert, im ersteren Falle besitzen sie unmittelbar einen Kontext –, so lassen sie sich doch problemlos ineinander überführen. Für beide Ansätze sprechen unterschiedliche Gründe, die sich häufig aus dem Anwendungskontext erklären lassen. URIs für jedes Tripel sind unmittelbar naheliegend, wenn einzelne Aussagen wie in einem Datenbanksystem häufig geändert werden, die Provenienz jedes einzelnen Eintrages festgehalten werden soll und dann jeweils situativ nach bestimmten Kriterien Aussagen zusammengefügt werden sollen. Dieses wird zum Beispiel in ResearchSpace [19] als Implementation einer *Linked Data Platform* (LDP) umgesetzt [143]. *named graphs* sind dann sinnvoll, wenn Aussagen zu einem bestimmten Zeitpunkt zusammengefasst werden sollen. Typische Beispiele hierfür sind Daten aus einer bestimmten Wissensbasis, bei denen der Kontext jedes einzelnen Datensatzes unerheblich ist. Wie *named graphs* für Provenienzaussagen genutzt werden können, findet sich in [38]. In unseren Beispielen haben wir es überwiegend mit dem zweiten Fall zu tun. In beiden Fällen lassen sich Eigenschaften wie Konsistenz und Widerspruchsfreiheit eingeschränkt auf Mengen von Graphen definieren.[20].

4.3 Kontexte und Graphen, Schließen mit *named graphs*

Nachdem in den vorhergehenden Kapiteln der Einsatz von *named graphs* zur Lösung des Versionierungs- und Provenienzproblems begründet wurde, soll der Ansatz in diesem Kapitel weiter ausgebaut und die Auswirkung ihres Einsatzes auf das Schließen über Graphen näher beleuchtet werden.

Die Modellierung in RDF bzw. OWL-DL erlaubt die Überprüfung der Daten auf Konsistenz und die Inferenz von Eigenschaften und Relationen mit Hilfe von Reasonern. Für die Anwendung in den Geisteswissenschaften ist jedoch eine Erweiterung der Modelle mit der Möglichkeit, Aussagen über Aussagen treffen zu können, notwendig. Ein Weg dazu ist der Einsatz von *named graphs* in der Verbindung mit Provenienzontologien, wie in Kapitel 4.5 besprochen werden wird. Aussagen über Aussagen verhindern jedoch zunächst den Einsatz von Reasonern, da sie in diesem System nicht mehr entscheidbar sind.

Im Folgenden soll jedoch argumentiert werden, dass die für die Forschung relevanten Aussagesysteme sich in geeigneter Weise einschränken lassen, so dass Schlussfolgerungen auch über erweiterte Systeme möglich sind.

4.3.1 Schließen mit *named graphs*

Sollen Aussagen in Sätzen über *named graphs* gebildet werden, muss aus dem Gesamtsystem eine Auswahl von Graphen gebildet und diese an den Reasoner übergeben werden. Über die eigentliche Frage der Konsistenz des dadurch entstehenden Gesamtgraphen hinaus können zusätzliche Fragen

[19] Siehe Abschnitt 6.6.
[20] Siehe 4.3.1.

gestellt werden, wie z.B. welche Teilmengen von Graphen konsistent sind, ob es einen maximalen konsistenten Graphen gibt oder wie sich solche Teilmengen konsistenter Graphen zueinander verhalten. Es geht also in diesem Kontext um graphentheoretische Fragestellungen der Topologie des Raumes konsistenter Graphen.

Für alle diese Fragen können die eigentlichen Aussagen in Form von Tripeln von den Kontexten getrennt beschrieben werden. Man hat also ein orthogonales Kreuzprodukt von $\{T\}\times\{C\}$. Geschlossen wird immer nur über die Teilmengen aus der Menge der Tripel. C beschreibt den Raum der Kontexte und ist eine abzählbare Menge.

Kontexte lassen sich in Subkontexte der folgenden Form aufteilen: $C = \{C_1, C_2\}$ wobei C_1 und C_2 so gewählt werden, dass C_1 alle Tripel beinhaltet, die widerspruchsfrei sind, und C_2 den Rest umfasst. Für unsere Fragen ist dann in erste Linie interessant, welche Kontexte widerspruchsfrei sind und ob es einen maximalen Kontext C_1 gibt, in dem möglichst viele Aussagen widerspruchsfrei zueinander bestehen können. Weitere Fragen sind, wieviele Cluster sich aus widerspruchsfreien Aussagen bilden lassen und inwieweit sich diese Cluster überlappen. Kann bestimmt werden, wie viele Aussagen aus einem Kontext herausgenommen werden müssen, damit dieser widerspruchsfrei wird oder können Aussagen gut begründet so mit Gewichten versehen werden, dass unwichtige Angaben unter Umständen vernachlässigt werden?

4.3.2 Unscharfes Schliessen, Handlungsräume und Überzeugungssysteme (Belief Systems)

Wir können an dieser Stelle nicht ausführlich auf die Probleme des Schließens eingehen. Für unseren Kontext ist jedoch wesentlich, dass der Ansatz, historische Daten modellierungstheoretisch zu erfassen, auch vor dem Hintergrund ungenauer, unvollständiger und teilweise widersprüchlicher Daten durchaus in einen erweiterten Kontext eingebettet werden kann, wenn eine formale Beschreibung dieser Aussagen gelingt. Vor diesem Hintergrund verweise ich auf die Arbeiten von Gerd Graßhoff, der am Beispiel der Entdeckung der Harnstoffsynthese zeigt, dass es mit Hilfe von Simulationen und einer dahinter liegenden Handlungslogik gelingen kann, wissenschaftliche Entdeckungen zu simulieren [97], sowie auf die Modellierung von Überzeugungssystemen (*belief systems*) im Kontext des AGM-Modells, das von Carlos Alchourrón, Peter Gärdenfors und David Makinson entwickelt wurde [86].[21]

4.4 CIDOC-CRM als ereignisbasierte Ontologie

Wie in 3.4 eingeführt, steht das Konzept des Ereignisses im Zentrum der folgenden Modellierung. Daher ist es naheliegend, auch ein Ontologiekonzept zur Grundlage zu nehmen, in dem Ereignisse den Ausgangspunkt für alle Beziehungen zwischen Objekten bilden. In diesem Kontext besitzen Objekte in unserem Modell keine unmittelbaren Eigenschaften, vielmehr sind diese lediglich Zuschreibungen, die als Ergebnis von Beobachtungen oder Annahmen zu einem bestimmten Zeitpunkt von

[21] Siehe auch Abschnitt 5.11.

Akteuren vorgenommen werden. Aus pragmatischen Gründen wird dies nicht immer auch unmittelbar so in das Modell eingehen. In unseren Modellen werden sich auch unmittelbare Aussagen wie **sr:has_name** oder **sr:worked_at** finden, da bei der Genese der in unser Modell eingehenden Daten die unmittelbaren Umstände der Datenerfassung nicht immer erfasst wurden. Es würden also Zuschreibungsereignisse entstehen, die letztendlich keinen Informationsgehalt haben, außer den, dass A von einer unbekannten Person an einem unbekannten Ort zu unbekannter Zeit eine Eigenschaft zugewiesen wurde. Unmittelbare Zuweisungen von Eigenschaften zu Objekten erfüllen in diesem Sinne in unseren Modellen immer die Funktion von Abkürzungen des vollen Pfades. Eine andere Form der Abkürzung ist die Benutzung des oben angeführten Konzeptes der *named graphs*. Wir fassen hier alle Zuschreibungen in einen *named graph* zusammen. Auch in diesem Falle ist dies dann semantisch einer Menge von Zuschreibungsereignissen gleichzusetzen.

Weitgehend konsequent ist der Ansatz von Zuschreibung und Abkürzungen im Rahmen des *Conceptual Reference Model*(CRM) des Internationalen *Komitees für Dokumentation* (CIDOC) des *International Council of Museums* (ICOM) umgesetzt. Ursprünglich für die komplexe Dokumentation von Objekten in Museen vorgesehen, ist dieses Modell mittlerweile auch in anderen Bereichen aufgenommen worden, in denen die Geschichte von Objekten sowie deren Eigenschaften beschrieben werden sollen, etwa in Form von *FRBRoo* [248] von den Bibliotheken oder innerhalb des Projektes Ariadne von der Archäologie. CRM ist auf den Webseiten des ICOM ausführlich dokumentiert, so dass eine Darstellung an dieser Stelle nicht notwendig ist. Besonders sei auf die graphischen Repräsentationen hingewiesen [84]. Diese geben einen guten Überblick darüber, wie Prozesse in CRM modelliert werden können. Die aktuellen Versionen von CRM liegen jeweils als ausführliche Textdokumente vor, in denen die Eigenschaften des Modell beschrieben werden. Es existieren darüberhinaus unterschiedliche Implementationen von CRM in formalen Sprachen. So existiert eine Implementation in RDFS, hier sind die aktuellen Versionen jeweils auf [84] zu finden. Erlangen-CRM bietet eine Implementation in OWL-DL1 [94] mit den aktuellen Versionen auf [216]. Erlangen-CRM ist auch der Ausgangspunkt für die in den Fallstudien vorgenommene Modellierung.

4.5 Provenienz und Versionierung mit *named graphs*

Für die Beurteilung der Qualität wissenschaftlicher Publikationen ist die Zuordnung von Aussagen zu ihren Urhebern wesentliche Voraussetzung. Hierbei geht es nicht in erster Linie um die Frage der Urheberschaft im Sinne des Urheberrechtes, sondern um die Anerkennung der wissenschaftlichen Leistung bei der Erstellung der Daten und die Möglichkeit für den Rezipienten, eine Aussage in ihren entsprechenden Entstehungskontext zu setzen.

Es zeichnet sich immer mehr ab, dass *named graphs* den Ausweg aus dem Problem der Provenienz von Aussagen in einem Triplestore aufzeigen. Verschiedene Ansätze sind zurzeit in Diskussion, wie diese genutzt werden können, um Graphen mit Aussagen über Autorschaft und Informationen zur Versionskontrolle anreichern zu können. Bei den in dieser Arbeit ausführlich dargestellten Beispielen werden *named graphs* systematisch eingeführt. Auch bei der Erstellung der Modelle wurden sie weitestgehend konsequent benutzt, um neue Versionen der im Laufe der Arbeit anfallenden Graphen im

Triplestore abzulegen und zu identifizieren. Dabei treten eine Reihe sehr konkreter praktischer Probleme auf, die zu einer in dieser Arbeit vorgeschlagenen und prototypisch vorgenommenen Erweiterung von SPARQL führen: nämlich der Möglichkeit, Abfragen über benannte Mengen von Graphen durchführen zu können (Abschnitt 4.5.1 und 4.5.2).

4.5.1 Versionsverwaltung von Graphen und Probleme des Schließens

Ein grundsätzliches Problem bei der Arbeit mit Daten, vor allem aber mit Datenbanken, ist das Problem der Versionierung. Generationen von Versionsverwaltungsprogrammen für Quellcode[22] und mittlerweile auch für Texte haben dazu geführt, dass hier ausgereifte Verfahren existieren, die Versionsvergleiche und Operationen wie das Verzweigen und Wiederzusammenführen von Versionen und Zweigen ermöglichen. Für Datenbanken stellt sich die Situation deutlich anders dar, und für Daten in Triplestores steht die Entwicklung noch weitgehend am Anfang. An Komplexität gewinnt dieses Problem noch dadurch, dass in einem System, das einen Wissensgraphen repräsentieren soll, neben Versionen im klassischen Sinne, die darauf zielen, vorhandene Daten schrittweise zu verbessern und weiterzuentwickeln, immer auch Versionen im Sinne von unterschiedlichen Interpretationen und Aussagen mitverwaltet werden müssen.[23] Umfasst ein solches System zusätzlich noch aus diesen Daten regelbasiert abgeleitete Daten, so müssen auch diese berücksichtigt und ihre Abhängigkeiten von den Versionen der Ursprungsgraphen protokolliert werden.

Es muss also nicht nur ein Weg gefunden werden, wie Versionen von Graphen verwaltet werden, sondern auch komplexe Mengen von Graphen beschrieben werden. Die Graphen, die sich in diesem Sinne weiterhin als konsistent erweisen, bilden die jeweils aktuelle Wissensbasis. Generische Anfragen an den Triplestore sollten sich in einer noch zu entwerfenden Erweiterung von SPARQL immer nur auf Aussagen in Graphen einer wohldefinierten Version der Wissensbasis beziehen, falls keine anderen weiteren Annahmen getroffen werden.

Im folgenden sei das Problem an einem einfachen Beispielen erläutert.

Graph 1 enthält die Aussagen:

```
<Bernd> <lebt in> <Köln> <graph 1>
```

In der neuen Version gilt nun aber:

```
<Bernd> <lebt in> <Brüssel> <graph 2>
```

Extern wird nun zusätzlich die Information

```
<Köln> <part_of> <Deutschland>
```

angenommen. In Version 1 gilt dann:

```
<Bernd> <lebt in> <Deutschland> <graph 1a_derived>
```

[22] Wie CVS [48], Mercurial [151] oder git [89].

[23] In der technischen Diskussion wird für diese Form von Versionen auch der Begriff „Variante“ benutzt. Siehe dazu [259].

In der neuen Version der Wissensbasis auf Grundlage von <graph 2> gilt diese Aussage nicht mehr. Diese Aussage muss bei der Änderung von <graph 2> in <graph 1> also auch korrigiert werden. Es müssen also bei Veränderungen von Teilaussagen jeweils auch alle abgeleiteten Aussagen entsprecht abgeändert werden. Dieses gilt zunächst unabhängig von der Frage, ob die Aussagen Revisionen sind, weil eine Aussage tatsächlich unzutreffend war oder weil Bernd zwischenzeitlich umgezogen ist.[24]

4.5.2 Verwaltung von Graphen

Im Beispiel der Datenbanken des Dombaus in Florenz (Abschnitt 9.2) werden aus den ursprünglichen Aussagen mithilfe von Algorithmen neue Aussagen (*statements*) erzeugt, es werden beispielsweise maschinenlesbare Datumsformate erzeugt oder sprachanalytische Methoden angewandt. (siehe 9.3).

Um hier sinnvolle Aussagen zu treffen, müssen wir jeweils Graphen, die sich aufeinander beziehen und in Abfragen gemeinsam verwendet werden sollten, immer zusammen behandeln..

Auch hier ein Beispiel: In der Version 1 lautet das Tripel

```
<http://FOO/12345> duomo:was_recorded_at "1420 giugno 15"
```

und daraus abgeleitet dann als maschinenlesbares Datum

```
<http://FOO/12345> duomo:has_XSD_date "1420-06-15"
```

Da wir die originären Daten und abgeleiteten Daten in unterschiedlichen Graphen halten wollten, befinden sich die beiden Tripel in unterschiedlichen Graphen.

Stellt sich nun in Version 2 des Graphen heraus, dass das Datum in 1422 geändert werden muss, so gilt also nun:

```
<http://FOO/12345> duomo:was_recorded_at "1422 giugno 15"
```

und entsprechend in Version 2 des Graphen der abgeleiteten Graphen

```
<http://FOO/12345> duomo:has_XSD_date "1422-06-15"
```

Berücksichtigt man nun nicht, dass sich beide Daten geändert haben, erhält man bei einer SPARQL-Abfrage ohne Berücksichtigung der Versionen.

```
<http://FOO/12345> duomo:has_XSD_date "1420-06-15"
<http://FOO/12345> duomo:was_recorded_at "1422 giugno 15"
<http://FOO/12345> duomo:was_recorded_at "1420 giugno 15"
<http://FOO/12345> duomo:has_XSD_date "1422-06-15"
```

Die richtigen Antworten erhält man also nur, wenn die Abfrage ausschließlich über die 2. Version erfolgt.

[24] Eine Lösung des Problems durch Einführung eines Konzeptes der Farbgebung von Tripeln als Erweiterung des Named-Graph-Konzeptes diskutieren Giorgos Flouris und andere in [82].

4.5.3 Mengen von Graphen und Versionierung

Um das Problem zu lösen, führen wir Mengen von Graphen ein. Diese lassen sich im Triplestore direkt verwalten. Hierbei benutzen wir das im Rahmen der Schaffung eines einheitlichen Vokabulars für *named graphs* vorgeschlagene Vokabular RDFG [144] und eine zusätzliche Klasse **graphSet**.

```
prefix rdfg <http://www.w3.org/2004/03/trix/rdfg-1>
prefix sr: <http://ontologies.mpiwg-berlin.mpg.de/scholarlyRelations/>

<http://set1> a sr:graphSet.
<http://graph1> a rdfg:Graph.
<http://graph2> a rdfg:Graph.
<http://graph1> rdfg:subGraphOf <http://set1>.
<http://graph2> rdfg:subGraphOf <http://set1>.
```

Anstelle einer Verwaltung im Triplestore ist auch ein Modell denkbar, in dem die einzelnen Graphen getrennt verwaltet und versioniert werden. In einem Filesystem bzw. Versionierungssystem würden dann die alten Versionen als RDF-Exporte gesichert, während der Triplestore immer nur die aktuellen Fassungen behält. Dies ist jedoch nur dann sinnvoll, wenn es sich um ein echtes Korrekturproblem handelt und die veraltete Version tatsächlich nur im Notfall als Referenz herangezogen werden soll. Eine weitere Möglichkeit wäre es, nur Unterschiede zwischen Versionen im Triplestore zu behalten, jedoch existieren auch dafür noch keine ausgereiften Realisierungen. Für einzelne Triplestore-Systeme existieren proprietäre Erweiterungen, die Abfragen über Gruppen vom Graphen (Virtuoso) [251] oder virtuelle Graphen (Blazegraph) [250] erlauben. Eine Standarisierung wäre hier wünschenswert.

4.6 Vorgehensweise bei der Umsetzung

In den Fallbeispielen werden wir sehen, wie die Umsetzung von historischen Fragestellungen in informationstheoretische Verfahren im konkreten Beispiel erfolgt ist. In dieser Arbeit soll verdeutlicht werden, wie die Interaktion zwischen historischer Forschung und Informatik gestaltet sein könnte, um neue Einsichten für beide Disziplinen zu erreichen. In den meisten Fallbeispielen ist die Bearbeitung noch nicht abgeschlossen und die vorgenommenen historischen Interpretationen sind daher noch vorläufig.

Die Vorgehensweise folgt in allen Beispielen einem gleichen Muster. Zunächst versuchen wir uns gemeinsam einen Überblick über die Datenlage zu verschaffen und suchen nach Heuristiken, die es uns ermöglichen, die unterschiedlichen Datenbestände auszuwerten. Für datenbankförmige Daten ist häufig der erste Schritt Namen und Personen eindeutig über unterschiedliche Beständen hinweg zu identifizieren. Für Volltexte erfolgt der Einsatz von OCR und Text-Mining-Verfahren, hier insbesondere *named entity recognition*, ebenfalls mit dem Ziel, zunächst Zusammenhänge zwischen den unterschiedlichen Daten aufzufinden.

Danach entwickeln wir ein erstes Datenmodell. In dieser Arbeit ist dieses paradigmatisch immer angelehnt an *CRM*. Die vorhandenen Daten werden dann in dieses Modell überführt. Auf die Daten wenden wir erste einfache Algorithmen an, um ein Verständnis dafür zu gewinnen, ob die Daten sich überhaupt in die Richtung der Forschungsfrage auswerten lassen. Diese Phase beinhaltet das Experimentieren mit unterschiedlichen Verfahren und Methoden, ist begleitet von einer Reihe von Rückschlägen, falschen Ergebnissen und manchmal auch voreiligen Schlüssen. Diese Phase erlaubt es jedoch, die Anforderungen an die Datenbestände stärker zu spezifizieren. Dies gilt einerseits für die Datenformate und Strukturen, und andererseits zeigt es auch, welche Daten noch erhoben werden müssen. Neue Daten erlauben dann weitere Iterationen. Pragmatisch heißt das immer wieder, leichte Anpassungen an den Algorithmen vorzunehmen, die wir in der Regel in Pythonnotizbüchern.[25] festgehalten haben.

[25] Dieses sind interaktive browserbasierte Skripte, die in unterschiedlichen Programmiersprachen geschrieben werden können. Ausführlicher werden wir diese in 6.1 einführen.

Kapitel 5

Modellierung und Netzwerkforschung

Die soziale Netzwerktheorie und die mathematische Graphentheorie, die dieser zugrunde liegt, wird in Lehrbüchern umfangreich und übersichtlich eingeführt.[1] Daher gibt dieses Kapitel keine grundlegende Einführung in die Netzwerktheorie. Dieses Kapitel will vielmehr verschiedene Ansätze aus unterschiedlichen Teilbereichen der Netzwerkanalyse zusammenführen, die in den folgenden Fallstudien eingesetzt wurden. Das Kapitel teilt sich daher in drei Teile auf: Im ersten Teil führen wir in die Grundbegriffe der Netzwerktheorie ein, danach in einige grundlegende Größen zur Beschreibung der Netzwerke und schließlich gehen wir auf einige spezifische Methoden ein, die zur Auswertung unserer Beispiele benutzt wurden. Letztere sind insbesondere Methoden zur Beschreibung und Analyse sich dynamisch entwickelnder Netzwerke.

5.1 Bemerkungen zu Grundbegriffen der Statistik

Wie auch in den anderen Teilen dieses Kapitels sollen hier nur die Begriffe eingeführt werden, die wir in den folgenden Abschnitten verwenden werden. Ohne eine solide Grundlage in der Anwendung der mathematischen Statistik sind die Modelle, die zur Bewertung von Netzwerken eingesetzt werden, nicht nachvollziehbar, so dass jedem, der den Einstieg in die Netzwerktheorie sucht, zumindest eine Einführung nahegelegt sei. Für die Modellierung von Netzwerken benötigen wir zusätzlich Grundlagen der Wahrscheinlichkeitstheorie sowie insbesondere ein Verständnis von Markov-Ketten und Monte-Carlo-Simulationen. Dafür sei auf die Standardliteratur verwiesen.[2]

5.2 Beschreibungsformen von Netzwerken - Matrizen und Graphen

Die folgende Zusammenfassung folgt im Wesentlichen [120] und [121].

Ein *Graph* $(\mathrm{N}, \mathfrak{g})$ besteht aus einer Menge von Knoten $\mathrm{N} = \{1, 2, \ldots, \mathrm{n}\}$ und einer Menge von Relationen $\mathfrak{g}$, die die Knoten in Beziehung setzen. Hierbei folgen wir im weiteren Verlauf der Matrix-

[1] Siehe zum Beispiel die Einführungen in die Netzwerktheorie von Dorothea Jansen [121] und Jackson [120]. Übersichtliche Kurzfassungen finden sich in [6], sowie [162] oder [14].

[2] Für eine Einführung in die Statistik siehe z.B. [155], zu Markov-Ketten [45],zu Markov-Ketten und Monte-Carlo-Simulationen [95].

darstellung. $\mathfrak{g} = ((g_{ij}))$ ist eine $n \times n$-Matrix mit $n \in \mathbb{N}$. $g_{ij} = 0$ bedeutet hier, dass die Knoten i und j nicht verbunden sind und $g_{i,j} \neq 0$, dass eine Verbindung besteht. Diese Matrix wird als *Adjazenzmatrix* des Graphen bezeichnet.

In einem *ungewichteten Graphen* gilt $g_{ij} \in \{0,1\} \forall i,j \in \{1,2,\ldots,n\}$ und in einem *gewichteten Graphen* $g \in \mathbb{R} \forall\, i,j \in \{1,..,n\}$. Eine Graph besitzt eine *Schleife*, wenn $g^{ii} \neq 0$ für ein $i \in \{1,..,n\}$. Ein Graph ist *schleifenfrei*, wenn er keine Schleife besitzt.

Reale Graphen, wie wir sie in unseren Fallbeispielen betrachten, bestehen in der Regel aus Beziehungen unterschiedlichen Typs. In diesem Fall existieren für jeden Typ Matrizen $\mathfrak{g}^1, \ldots, \mathfrak{g}^m$, die die Beziehungen zwischen den einzelnen Knoten für jeden Typ beschreiben. Wir haben damit für den Graphen die Darstellung $(N, \{\mathfrak{g}^1, \ldots, \mathfrak{g}^m\})$. Interessiert nur die Beziehung der Knoten in dem Graphen unabhängig von der Art der Beziehung so erhalten wir $(G, \bar{\mathfrak{g}})$ mit $\bar{\mathfrak{g}} = \frac{1}{m} \sum_{k=1}^{m} \mathfrak{g}^m$. Die Summe der Matrizen beschreibt dann den gesamten Graphen.

In einem *ungerichteten Graphen* gilt $g_i^j = g_j^i, \forall i,j \in \{1,\ldots,n\}$ anderenfalls ist der Graph gerichtet. Graphen lassen sich stets in gerichtete und (vollständig) ungerichtete Anteile zerlegen.

Das heißt also eine symmetrische Matrix repräsentiert einen ungerichteten Graphen. Ein schleifenfreier Graph ist durch eine Matrix mit 0 auf der Hauptdiagonalen repräsentiert.

5.2.1 Bipartite und monomodale Graphen

Wir unterscheiden zwei Grundtypen von Netzwerken. Ein Graph ist *monomodal*, wenn er nur einen Typ von Knoten besitzt. Ein *bipartiter* Graph besitzt genau zwei Typen von Knoten. Diese sind dadurch ausgezeichnet, dass alle Kanten nur jeweils zwischen zwei verschiedenen Typen existieren. Daraus folgt, dass die Repräsentation eines monomodalen Graphen immer eine quadratische Matrix ist, während ein bipartiter Graph auch durch andere Matrizen repräsentiert sein kann. Auf komplexeren Strukturen mit mehr als zwei Knotentypen und unterschiedlichen Kanten kommen wir in 5.8 zu sprechen. Letztendlich spielen diese die wesentliche Rolle bei der Beschreibung der Strukturen, die wir in den Fallbeispielen untersuchen werden. Jedoch sind diese ohne grundlegende Begriffe aus der Theorie der bipartiten und monomodalen Graphen nicht zu beschreiben.

Bipartite Projektion

Ein ungerichteter bipartiter Graph lässt sich immer auf natürliche Art und Weise auf zwei monomodale Graphen abbilden, in dem jeweils zwei benachbarte Kanten des bipartiten Graphen auf eine Kante des monomodalen Graphen abgebildet werden. Es entsteht dann jeweils ein monomodaler Graph aus einer der Kantensorten. Erlaubt man für den projizierten Graphen, dass er Schleifen besitzt, so werden einzelne Kanten auf Schleifen abgebildet. Bei gerichteten Graphen existiert keine natürliche Projektion. Gleiches gilt für gewichtete Graphen. In unseren Fällen projizieren wir ausschließlich auf ungerichtete ungewichtete Graphen.

5.3 Charakteristische Größen in Netzwerken

Aus der Topologie der Graphen ergeben sich eine Reihe von charakteristischen Größen. Diese sind in der sozialen Netzwerktheorie in der Regel mit Interpretationen über das Verhalten und die Stellung von Akteuren als Knoten in den Netzwerken, die durch die entsprechenden Graphen repräsentiert werden, verbunden. Diese Größen werden allgemein als Zentralitäten bezeichnet.

Bei den folgenden Darstellungen benutze ich zur Vereinfachung in der Regel die englischen Bezeichnungen für die entsprechenden Größen, da sie bei weitem häufiger eingesetzt werden als die entsprechenden deutschen Ausdrücke.

5.3.1 Degree (Grad)

Der *degree* eines Knotens bezeichnet die Anzahl der mit ihm verbundenen Kanten. In einem gerichteten Graphen wird hierbei *in-degree* und *out-degree* unterschieden. In einem gewichteten Graphen lässt sich entsprechend ein *degree* unter Berücksichtigung der Gewichte definieren, in dem die Gewichte aller Kante aufaddiert werden. Wir betrachten in der Regel jedoch immer nur ungewichtete *degrees*.

In der Matrix-Darstellung $A = ((a_{ij}))$ eines ungerichteten Graphen mit n Knoten ist der *degree* des k-ten Knotens die Summe einer Zeile (bzw. Spalte):

$$\mathrm{degree}(k, A) = \sum_i a_{ki}$$

Bei einem gerichteten Graphen gilt:

$$\mathrm{degree}_{\mathrm{in}}(k, A) = \sum_i a_{ki}$$

und

$$\mathrm{degree}_{\mathrm{out}}(k, A) = \sum_j a_{jk}$$

mit $j, i \in \{1, \ldots, n\}$.

5.3.2 Kürzester Pfad, Zusammenhang, Radius, Komponenten und Betweenness

Der *kürzeste Pfad* zwischen zwei Knoten k und l eines Graphen ist der Teilgraph mit dem Anfangspunkt k und Endpunkt l, dessen Kantenanzahl minimal ist. Kürzeste Pfade sind schleifenfrei, aber nicht notwendig eindeutig. Ein ungerichteter Graph ist *zusammenhängend*, wenn zwischen allen Knoten des Netzwerkes ein kürzester Weg existiert. Der *Radius* eines zusammenhängenden Graphen ist die Anzahl der Kanten des längsten kürzesten Pfades in einem Graphen. In einem gerichteten Graphen ist der kürzeste Pfad von k nach l der kürzeste vollständig gerichtete Teilgraph. Offensichtlich muss dieser, auch wenn der ungerichtete Graph zusammenhängend ist, nicht existieren. Ein Graph ist *gerichtet zusammenhängend*, wenn ein gerichteter Teilgraph zwischen allen Paaren von Knoten

des Graphen existiert. In der Literatur wird ein solcher Graph auch als *stark* zusammenhängend bezeichnet bzw. als *schwach* zusammenhängend, wenn lediglich der ungerichtete Graph zusammenhängend ist. Die zusammenhängenden Teilmengen des Graphen bezeichnen wir als *Komponenten*. Aus den kürzesten Pfaden wird der Begriff der *betweenness* eines Knoten abgeleitet. Dieses ist die Anzahl aller kürzesten Pfade, die durch einen Punkt gehen, geteilt durch die Anzahl aller kürzesten Pfade eines Netzwerkes. Diese Größe skaliert mit der Anzahl der Paare von Knoten. Wir betrachten daher später insbesondere beim Vergleich der zeitlichen Entwicklung eine normierte Form der *betweenness*, indem wir durch die Anzahl der Paare ohne den Knoten selbst dividieren, also durch $(N-1)(N-2)/2$.

Die *betweenness* gibt somit im weitesten Sinne Auskunft über die Möglichkeit, Informationsflüsse zu kontrollieren, Einfluss auszuüben, bzw. über das Potential der Informationsagglomeration an solchen Stellen.[3]

5.3.3 Closeness

Die *closeness* (Nähe) eines Knotens beschreibt umgekehrt, wie nah ein Knoten in einem Netzwerk allen anderen Knoten ist. Bei nicht zusammenhängenden Knoten ist dieses nicht unmittelbar zu definieren. Es gibt hier eine Reihe von Möglichkeiten dafür, eine Entfernung anzusetzen. Wir betrachten in unseren Fällen in der Regel lediglich Nähebeziehungen innerhalb einer Komponente, da wir diese als Maßstab für die innere Struktur dieser Komponenten verstehen.

Als Maß für die Nähe wird auch hier wieder auf die kürzesten Pfade zurückgegriffen und wir definieren den Kehrwert der Summe der Länge aller kürzesten Pfade von einem Knoten zu allen anderen als die Nähe eines Knotens. Die maximale Zentralität ist dann in einem Graphen von n Knoten $1-n$. Diesen Faktor nutzen wir, um diesen Kehrwert zu normieren. Es gilt also für den i-ten Knoten in unserem Graphen mit d_{ij} für die Länge des kürzesten Pfades von i nach j:

$$C(i) = \frac{n-1}{\left(\sum_j d_{ij}\right)}$$

5.3.4 Zentralisierung des Graphen

Um Netzwerkstrukturen zu vergleichen, stellt sich die Frage nach aussagekräftigen Werten, die sich von Eigenschaften der Knoten auf den gesamten Graphen oder einen Teilgraphen übertragen lassen. Wir benutzen im Folgenden die auf Freemann zurückgehende Definition[121, S.139] für die *betweenness* des gesamten Graphen, wenn $B(i)$ die *betweenness* für den Knoten i bezeichnet:

$$C_B = 2\frac{\sum_i (\max(B) - B(i))}{(n-1)^2(n-2)}$$

[3] Zugleich ist diese Größe in unseren Kontexten kritisch, da sie sich gegenüber dem Hinzufügen und Entfernen von Knoten und Kanten instabil verhält. Insbesondere in unseren historischen Kontexten führt das Auffinden möglicher weiterer Verbindungen in einem Netzwerk häufig zu starken Ausschlägen der *Betweenness* bei Personen an besonders markanten Stellen, wenn diese, beispielsweise in internationalen Kontexten, die einzige bekannte Beziehung zwischen unterschiedlichen Seiten beschreiben. Informelle Kontakte, für die es keine historischen Belege gibt, führen so zu teilweise nur schwierig zu interpretierenden Entwicklungen.

Für die Closeness mit $C(i)$ *closeness* für jeden Punkt gilt entsprechend:

$$C_C = \frac{\sum_i (\max(C) - C(i))}{(n-1)(n-2)}(2n-3).$$

Die Faktoren normieren hierbei jeweils den Ausdruck, so dass dessen Wert zwischen 0 und 1 liegt.

5.3.5 Prestige

Neben der *betweenness* betrachten wir in unseren Netzwerken noch das Proximity-Prestige [121, S.139]. Dieses ist eine vom *degree* abgeleitete Größe. Anstelle lediglich die unmittelbaren Nachbarn zu berücksichtigen, betrachten wir zusätzlich deren Vernetzungsgrad im Netzwerk. Dadurch werden in stark hierarchisierten Netzwerken, die Personen an der Spitze der Hierarchie stärker bewertet. Anders als mit *closenness* und *betweenness* erhalten wir so eine Größe für die Hierarchisierung eines Netzwerkes.

Das Prestige eines Knotens j definieren wir damit (6.19 in [121, S.139]) als

$$P(j) = \frac{I_j/(n-1)}{\sum_j d_{ij}/I_j},$$

wobei I_j die Anzahl der von j erreichbaren Knoten angibt und d_{ij} die Länge des kürzesten Pfades von i nach j. Als Maß für die Hierarchisierung kann dann die Varianz des Prestiges betrachtet werde:

$$S_p = \sqrt{\sum_i \frac{\left(P(i) - \bar{P}\right)^2}{n}}$$

5.4 Substrukturen, Cluster

Wie wir in Abschnitt 5.13 beschreiben werden, besteht unser Interesse an der Anwendung der Netzwerktheorie insbesondere darin, Substrukturen und Ordnungsprinzipien, die diese Substrukturen bestimmen, charakterisieren zu können. Das Auffinden solcher Teilstrukturen ist eine der zentralen Herausforderungen der Graphen- und Netzwerktheorie. Dies gilt sowohl für die mathematische Seite als auch für die Seite der sozialwissenschaftlichen und historischen Interpretationen dieser Strukturen. Dieses Feld ist daher im Zentrum eines regen Austausches der unterschiedlichen Disziplinen und verschiedenste Ansätze zur Lösung dieses Problems wurden in den letzten Jahren und Jahrzehnten diskutiert. Unterschiedliche Methoden werden sowohl in den Lehrbüchern wie [120] oder [121] als auch in der aktuellen Literatur, siehe z.B. [161] oder [72], diskutiert. Die Diskussion wird getragen von einem Wechselspiel von sozialwissenschaftlicher Interpretation und algorithmischen Notwendigkeiten.[4]

[4] Hilfreich für eine Übersicht der unterschiedlichen Implementationen und Algorithmen zur Berechnung von Community-Strukturen und Clustering sind insbesondere auch die zumeist ausführlichen Dokumentationen der unterschiedlchen Softwarepakete zur Netzwerkanalyse wie igraph [118] oder Networkx [160], die sowohl auf Beispielanwendungen als auchauf die dahinter stehende Literatur verweisen. Siehe auch Abschnitt 6.2.

In zwei unserer Fallbeispiele steht die Frage nach der Bildung von Substrukturen und deren Rolle im Zentrum. Wir fragen uns, welche Faktoren zu ihrer Bildung geführt haben und ob sich Eigenschaften in den Netzwerkstrukturen wiederfinden, von denen wir in der historischen Analyse annehmen, dass sie zur Formierung eines Teilsystems einer handelnden Gruppe geführt haben. Im Sinne unseres Ansatzes einer netzwerktheoretischen Beschreibung als Heuristik für die historische Forschung ist die Beantwortung dieser Frage in der Regel mit mehreren Problemkomplexen verbunden. Das größte Problem ist hier stets die Datenbasis und die daraus abgeleitete Interpretation. Wir können im Allgemeinen nicht von einem vollständigen Datensatz ausgehen, und netzwerktheoretische Ergebnisse bei der Bestimmung von Clustereigenschaften können stark variieren, wenn sich das Netzwerk auch nur geringfügig ändert. Weiterhin ist die Interpretation von Beziehungen und deren Gewichtung in der Regel eine Herausforderung für die historische Forschung, daher ist das Gewicht einer Kante oftmals zunächst ein freier Parameter, für den plausible Größenannahmen erst durch die Netzwerkanalysen gefunden werden können. Wir werden dies insbesondere bei den Relationen sehen, die auf Grund gemeinsamer Zugehörigkeit zu Institutionen entstehen. Wir versuchen dieses Problem insoweit einzugrenzen, dass wir die Veränderungen der Ergebnisse abhängig von Variationen äußerer Parameter systematisch analysieren. Daher liegt ein Schwerpunkt dieser Arbeit insbesondere auf Hilfsmitteln, um diese Analysen möglichst interaktiv zur Verfügung zu stellen.

In der sozialen Netzwerktheorie werden solche Probleme ebenfalls angegangen. Wir grenzen uns hierbei insofern von sozialwissenschaftlicher Methodik ab, als dass wir immer auf sekundäre Daten angewiesen sind, um Interaktionsbeziehungen zu beurteilen. Unsere Quellen wurden selten mit der Absicht erstellt, Beziehungsgeflechte möglichst objektiv zu beurteilen. Die Umformung dieser Quellen in quasi historisch-soziologische Daten ist Forschungsarbeit, bei der zumindest die Gefahr besteht, die derzeitige Auffassung über die Stärke und Funktion sozialer Interaktionen stärker widerzuspiegeln als das tatsächliche historische Interaktionsnetz.[5]

Zur Bewertung von internen Strukturen wurde in der sozialen Netzwerkforschung eine Reihe von Konzepten eingeführt. Für eine detaillierte Einführung verweise ich auf das Kapitel 8 von Jansen [121].

5.4.1 Dichte

Die Dichte eines Netzwerkes ergibt sich durch das Verhältnis der vorhandenen Kanten m zu den möglichen Kanten eines Netzwerkes mit n Knoten. Wir haben also:

$$\mathrm{D}_{\text{gesamt}} = \frac{\mathrm{m}}{\frac{\mathrm{N}(\mathrm{N}-1)}{2}}.$$

Die Innendichte eines Teilnetzes mit z Knoten ergibt sich entsprechend durch das Verhältnis der tatsächlichen zu den möglichen Binnenkanten:

$$\mathrm{D}_{\text{binnen}} = \frac{\mathrm{i}}{\frac{z(z-1)}{2}}.$$

[5] Siehe dazu auch Abschnitt 5.12.

Entsprechend ist die Außendichte durch die Anzahl a der Kanten des Netzes, die Mitglieder des Teilnetzes mit externen Knoten verbinden, bestimmt:[6]

$$D_{\text{außen}} = \frac{a}{z(N-z)}.$$

Einen Sozialen Kreis erhalten wir dann, wenn die Innendichte (signifikant) größer als die Dichte des Gesamtnetzes und diese wiederum größer als die Außendichte ist.

5.4.2 Clusteringkoeffizient

Der Clusteringkoeffizient beschreibt die Kompaktheit eine Netzwerkes und wird zurückgehend auf Watts und Strogatz [258] als:[7]

$$C = \frac{3 \times (\text{Anzahl der Dreiecke in einem Graphen})}{(\text{Anzahl der Triaden })} \tag{5.1}$$

Im Paket igraph für R und für Python entspricht dieses der globalen Transitivität [242].

5.4.3 Modularität und Clustering

Die Frage nach Clustern bzw. Communities in Netzwerken ist eng mit zwei Problemen verbunden: Erstens, wie misst man die Nähe zweier Knoten in einem Netzwerk und zweitens, wie lassen sich aus der individuellen Nähebeziehung globale Strukturen des Graphen ableiten? Beide Probleme lassen sich letztendlich nicht losgelöst von der historischen Fragestellung angehen. Unabhängig davon ist jedoch zusätzlich ein Kriterium notwendig, das es ermöglicht, die Ergebnisse von Cluster- bzw. Communitybildung zu bewerten. Ein Ansatz dazu beruht auf dem Konzept der Modularität, das Newman in einer Reihe von Artikeln [163, 47, 161] entwickelt.

Die Modularität (modularity) Q bestimmt, wie sehr Knoten eines Netzwerks mit einer bestimmten Eigenschaft dazu tendieren, enger verbunden zu sein, d.h. also mehr Kanten ausbilden, als dies bei einem zufälligen Graphen der Fall wäre. Damit kommt man zu einer allgemeinen Definition der Modularität [161, Gl. 17]:

$$Q = (\text{number of edges within communities}) - (\text{expected number of such edges}).$$

Hierbei ist die erwartete Zahl mit dem Problem verbunden, ein möglichst nachvollziehbares Vergleichsmodell zu definieren. Eine Setzung dafür ist die Annahme einer zufälligen Wahrscheinlichkeit bei Beibehaltung der *degree*verteilung [47].

$$Q = \frac{1}{2m} \sum \left[a_{ij} - \frac{\text{degree}(i) * \text{degree}(j)}{2m} \right] \delta(c_i, c_j),$$

mit m Anzahl der Kanten des Graphen, c_x der Größe einer Eigenschaft eines Knotens x, $\delta(x, y)$ ist 1 wenn $x = y$, sonst 0.

Führen wir die Laplace-Matrix $\mathbb{L}$ des Graphen ein

[6] Der Faktor 2 entfällt hier, da es hier nicht zu Doppelzählungen kommt.

[7] Für eine Motivation dieser Form der Darstellung siehe auch [17].

$$\mathbb{L}_{ij} = \frac{\text{degree}(i)}{2m}\delta(i,j) - a_{ij},$$

so erhalten wir[8]

$$Q = \frac{1}{2m}s_i s_j \mathbb{L}_{ij} \text{ mit } s_i = \begin{cases} +1 & \text{wenn der Knoten i die Eigenschaft hat} \\ -1 & \text{sonst} \end{cases},$$

bzw. in Matrizendarstellung:

$$Q = \frac{1}{2m}s^T \mathbb{L} s.$$

Analog lässt sich die Modularität für eine Aufteilung in k-Communities definieren [163, Gl.5]:

$$Q_k = \sum_i \left(e_{ii} - a_i^2\right),$$

wobei $a_i = \sum_j e_{ij}\ e_{ij}$ hierbei die symmetrische $k \times k$-Matrix ist, die den Anteil der Kanten beschreibt, die die i-te Community mit der j-ten verbindet.

5.4.4 Clustering-Algorithmen

In den Fallbeispielen werden wir sehen, dass Algorithmen, die Häufungen von Knoten in einem Netzwerk identifizieren, ein wesentliches heuristisches Hilfsmittel für unsere Untersuchungen sein werden. Um diese Methoden anzuwenden, ist es jedoch unerlässlich, zumindest in Grundzügen, die hinter den unterschiedlichen Clusteralgorithmen stehenden soziologischen bzw. strukturellen Annahmen zu verstehen. An dieser Stelle werden wir die mathematische Herleitung nicht im Detail beschreiben, sondern im Wesentlichen auf die Literatur verweisen.

> If we are concerned with the process that generated the network in the first place, we should use methods based on some underlying stochastic model of network formation.To study the formation process, we can, for example, use modularity [163], mixture models at two or more [46] levels, Bayesian inference [112], or our cluster-based compression approach [211] to resolve community structure in undirected and unweighted networks. If instead we want to infer system behavior from network structure, we should focus on how the structure of the extant network constrains the dynamics that can occur on that network. To capture how local interactions induce a system-wide flow that connects the system, we need to simplify and highlight the underlying network structure with respect to how the links drive this flow across the network.[9]

betweenness-basierte Verfahren

Clustering aufbauend auf *betweenness* wird von M.E.J. Newman und M. Girvan in [163, Chapter II] ausführlich beschrieben. Allgemein gilt, dass *betweenness*-Zentralitäten immer die Bedeutung einer

[8] Für die Ableitung siehe [161, sec II].

[9] Zitat von [210, P. 14], Referenzen übernommen aus dem Zitat.

Kante oder eines Knotens für Austauschprozesse in einem Netzwerk beschreiben. Hierbei sind verschiedene Austauschprozesse denkbar. Oben haben wir schon die pfadbasierte *betweenness* eingeführt[10], die als ein Maß für Informationskontrolle dienen kann. Eine Alternative dazu sind flussbasierte Modelle. Diese beruhen auf der Bestimmung der Menge einer noch zu definierenden Ladung, die über einen Kante oder einen Knoten fließt. Clustering auf der Grundlage von unterschiedlichen Ansätzen für die *betweenness* folgt in allen Fällen dem von Newman und Girvan beschriebenen Algorithmus einer schrittweisen Aufteilung des Netzwerkes durch das Entfernen der Kanten mit jeweils der höchsten *betweenness*. Vorausgesetzt, das Netzwerk zerfällt dadurch in mehrere nicht mehr zusammenhängende Bestandteile, ergibt sich so ein hierarchisches Clustermodell.

5.4.5 Clustering über die Eigenwerte der Adjazenzmatrix

Ein anderes Verfahren der Clusterbildung erhalten wir beim Verfahren, das sich an den Eigenwerten der Adjazenzmatrix orientiert. Die Eigenräume der Adjazenzmatrix liefern Aussagen über die Struktur des Netzwerkes in der Hinsicht, welchen Einfluss die Ähnlichkeit benachbarter Knoten auf die Bildung von Beziehungen gespielt hat (Freunde von Freunden werden Freunde). Das Verfahren zur Bildung von Clustern über die Eigenwerte der Adjazenzmatrix wird ausführlich z. B. in [253] beschrieben und besteht im Wesentlichen darin die k-ersten Eigenräume zu bestimmen und diese als Ausgangspunkte für ein k-Means-Clustering zu benutzen.

5.4.6 Infomap

Von besonderer Relevanz besonders für den Einsatz im Fall von Kozitationsbeziehungen[11] und später für die Untersuchung der Koinfluenzbeziehungen[12] ist das Clustering für Netzwerke, das auf der sogenannten *Map Equation* beruht [210, 211]. Hier ist die Idee, das Netzwerk aus der Sicht eines Random Walkers in dem Netz zu klassifizieren. Der Mechanismus ist direkt auch auf Multilevel-Netzwerke anwendbar. Auf der Webseite von Daniel Edler und Martin Rosvall (www.mapequation.org) finden sich instruktive Anwendungsbeispiele und frei verfügbare Software für die Anwendung des Algorithmus. Im Wesentlichen wird ein Cluster hier über die Aufenthaltswahrscheinlichkeit eines *random walkers* bestimmt.

5.4.7 Blockmodelle

Blockmodelle sind eine weitere Möglichkeit zur Klassifizierung von Knoten in Netzwerken. Die zentrale Idee ist es hier, Gruppen von Knoten zu finden, die sich bezüglich ihrer Kanten möglichst ähnlich verhalten, die also in diesem Sinne austauschbar sind. Eine detaillierte Beschreibung der unterschiedlichen Verfahren zur Bestimmung von Blockmodellen unter verschiedenen Randbedingungen und Annahmen für die Ähnlichkeit findet sich in [171]. Auch für dieses Verfahren gilt, dass eine Erweiterung auf Multilevel l Netzwerke möglich ist [276].

[10] Siehe Abschnitt 5.3.2.
[11] Siehe Abschnitt 7.7.
[12] Siehe Abschnitt 8.8.

5.5 Dynamische Entwicklung von Graphen und ihre Modellierung

Was bedeutet die Anwendung von Modellierungsverfahren in unseren Kontext? Ziel der Experimente mit Modellierungsansätzen ist nicht in erster Linie die Vorhersage von Parametern wie in der klassischen sozialwissenschaftlichen Forschung – wir verstehenden die Modelle vorrangig als Hinweise auf historische Mechanismen. Es geht es uns darum zu untersuchen, ob die soziologischen und ökonomischen Annahmen, die hinter den Modellen stehen, die den historischen Sachverhalt in gewisser Genauigkeit simulieren können, auch Hinweise auf die Ursachen historischer Entwicklung geben können. Auch hier sind rechnergestützte Methoden heuristische Instrumente und damit Teil eines hermeneutischen Prozesses. Häufiger als positive Übereinstimmung können in der praktischen Arbeit keine übereinstimmenden bzw. konvergierenden Modelle gefunden werden, womit auch die hinter diesen Modellen liegenden Erklärungsmuster in Frage gestellt werden. Auf diese Art und Weise erwarten wir von der Zusammenarbeit von historischer Netzwerkforschung mit aktuellen Modellbildungen auch einen Impuls für die Entwicklung aktueller Verfahren. In Abschnitt 5.12 werden wir ausführlicher auf die Risiken eingehen, zunächst jedoch eine kurze Übersicht über die Ansätze geben, die bei der Analyse der Netzwerke verfolgt wurden, sowie über einige zentrale Grundbegriffe aus der Theorie der dynamischen Entwicklung.

5.5.1 Power Laws und Skalenfreiheit

Von besonderem Interesse bei der Analyse von Netzwerken, bzw. allgemeiner dem Verhalten von größeren Mengen, sind Zusammenhänge, bei denen relative Veränderung nur von der Veränderung eines Parameters abhängt und zugleich nicht von der Größe der Ausgangsgröße abhängen. Solche Veränderungen bezeichnet man als *skalenfrei* oder *skalenunabhängig*. Wie sich leicht zeigen läßt, wird dieses von Abhängigkeiten erfüllt, die die folgende Form haben.

$$f(x) = cx^k + b$$

Auch alle Linearkombinationen für Funktionen mit unterschiedlichem k erfüllen die Bedingung der Skalenfreiheit. Es gilt also:

$$f(x) = \sum_{k=1..n} c_k x^k \, \forall k \in \mathbb{N}$$

Abhängigkeiten, die dieser Gleichung folgen, werden als *power laws* oder *Potenzgesetze* bezeichnet,

5.5.2 Zufällige Graphen

Eine der wesentlichen Methoden zur Analyse von realen Netzwerken ist der Vergleich ihrer Eigenschaften und ihrer dynamischen Entwicklung mit dem Verhalten von zufälligen Netzwerken. Grundlegende Abweichungen vom Verhalten zufälliger Graphen geben Hinweise auf besondere Ordnungsmechanismen. Zum Verständnis werden neben der Graphentheorie Grundlagen der Wahrscheinlichkeitstheorie benötigt.[13]

[13] Für eine grundlegende Einführung in die Wahrscheinlichkeitstheorie siehe die zahlreiche Lehrbuchliteratur z.B. [132, 185].

Das Verhalten zufälliger Graphen ist seit den ersten grundlegenden Arbeiten von Erdős und Rényi [73] in den späten 1950er und 1960er Jahren systematisch untersucht worden. Matthew O. Jacksons ausführliches Lehrbuch *Social and Economic Networks* [120] gibt eine umfassende Einführung in die Techniken und Methoden zur Modellierung von Netzwerken.

Wir werden in den Fallbeispielen Modellierungsansätze im Wesentlichen einsetzen, um Hinweise darauf zu bekommen, wie stark einzelne Parameter die Struktur unser Netzwerke beeinflussen. Dazu vergleichen wir die Netzwerke, die sich aus der historischen Analyse ergeben, mit zufälligen Netzwerken. Wir wollen dabei sehen, wie ähnlich unsere Netzwerke diesen zufälligen Netzwerken werden, deren Entstehungsbedingungen wir kontrollieren können. Dazu betrachten wir einerseits zufällige Graphen nach dem Modell von Erdős und Rényi, und andererseits folgen wir dem Barabási-Albert-Modell bzw. Newman, Strogatz und Watts.

Modelle nach Erdős und Rényi folgen im Wesentlichen einem einfachen Prinzip: Es wird eine feste Anzahl von Knoten und eine feste Wahrscheinlichkeit p für die Bildung eines Knotens bzw. äquivalent dazu eine feste Anzahl von Kanten vorgegeben. Untersucht wird dann die Struktur der Netzwerke, die sich daraus bei entsprechender zufälliger Bildung des Netzwerkes ergeben [120, S. 1.2.3]. Die wesentliche Eigenschaft der Netzwerke, die sich in dieser Form bilden, ist, dass ihre *degree*-Verteilung $\mathrm{d}(\mathrm{n})$ für den Grenzfall großer Netzwerke einer Poisson-Verteilung folgt. Es gilt also

$$\mathrm{d}(\mathrm{n}) = \mathrm{e}^{-(\mathrm{n}-1)\mathrm{p}}$$

5.5.3 Skalenfreie zufällige Graphen und hybride Modelle

Ausgehend von realen Graphen insbesondere von Kollaborationsnetzwerken untersuchen Newman, Strogatz und Watts [164] Netzwerke, die nicht durch die Poisson-Verteilung der Kanten vorgeben sind, sondern bei denen die Dichteverteilung vorgegeben ist. Ein Sonderfall in diesem Zusammenhang ist das von Barabási und Albert in [15] untersuchte Modell, in dem die Dichteverteilung durch ein *power law* vorgegeben wird. Alle weiteren Eigenschaften des Graphen sind dagegen zufällig verteilt. Barabási und Albert zeigen, dass diese Graphen sich ergeben, wenn man eine Eigenschaft, die sie als *preferential attachment* bezeichnen, annimmt. Im Wesentlichen heißt dieses, dass die Wahrscheinlichkeit, dass eine neue Bindung entsteht, proportional zu vorher bereits bestehenden Beziehungen ist. Jackson baut diese in [120, Kap 5.3.] zu hybriden Modellen aus. Hybride Modelle sind hierbei Modelle, die sich aus einer Kombination der zufälligen Wahl von Bindungen und dem *preferential attachment* ergeben. Insbesondere zeigt er dort [120, S.140] , dass sich in der Hauptfeldnäherung eine erzeugende Funktion für den *degree* in erster Näherung folgender Form ergibt:

$$\log(1 - \mathrm{F}(\mathrm{d})) = \frac{2}{1-\alpha} \log\left(\mathrm{m} + \frac{2\alpha\mathrm{m}}{1-\alpha}\right) - \frac{2}{1-\alpha} \log\left(\mathrm{d} + \frac{2\alpha\mathrm{m}}{1-\alpha}\right), \tag{5.2}$$

mit m dem halben mittleren *degree* sowie d entsprechend dem *degree*. α gibt das Verhältnis zwischen den beiden Formen der Wahl von Beziehungen an. $\alpha \to 1$ entspricht dem Extrem, dass die Wahl

vollständig zufällig ist, und der Fall $\alpha = 0$ bedeutet entsprechend die gegenteilige Annahme.[14]

5.5.4 *Small worlds*

Unter *small worlds* verstehen wir Netzwerke mit einer relativ zur Größe des gesamten Netzwerks kleinen typischen Pfadlänge. Charakteristisch für reale Netzwerke ist neben einer Small-World-Eigenschaft, dass der Clusteringkoeffizient groß ist, während dieser bei zufälligem Netz mit gleichem mittleren *degree* um Größenordnungen kleiner ist[15]

In [6, S68 ff.], wird das Verhalten von Small Worlds nach dem Modell von Watts und Strogatz [258] in der Erweiterung Newman and Watts [165] ausgiebig diskutiert. Abbildung 5.1 zeigt das Vorgehen von Watts und Strogatz.

Die Annahme sind hier ringförmige Netzwerke, die durch das Hinzufügen von „Abkürzungen“ zu *small worlds* werden [258, S.441].

Insbesondere ist folgendes Verhalten auffällig: Die typische Pfadlänge L wächst für kleine p linear, während sie nach einem Phasenübergang für große p dann nur noch logarithmisch wächst. Anschaulich lässt sich im Modell erläutern, dass mit steigendem p die Wahrscheinlichkeit steigt, dass Abkürzungen zwischen den einzelnen Regionen des Netzwerkes entstehen und damit die typische Weglänge mit jeder hinzugefügten Abkürzung deutlich sinkt. Für das Absinken existiert hierbei eine kritische Grenze $p_c = 2/NK$, ab der die typische Pfadlänge zu sinken beginnt, wobei N die Anzahl der Knoten und K die Anzahl der nächsten Nachbarn vorgibt. In diesen Graphen gilt außerdem für den Clusteringkoeffizienten:

$$C = \frac{3(K-2)}{4(K-1)} \tag{5.3}$$

Allgemein gilt: Es existiert eine von p abhängige Größe N^*, die bestimmt, wann sich das Verhalten ändert, so dass gilt: für $N < N^*$ $L \sim N$, für $N > N^*$ jedoch $L \sim \ln(N)$. Diese Größe wird aus der anschaulichen Interpretation, dass in diesem Moment Abkürzungen entstehen, als *cross-over size* bezeichnet.

Aus Modellrechnungen ergibt sich dann nach Gleichung 69 in (Albert und Barabási 2002) für die typische Länge L

$$L(N,p) \sim \frac{N^{1/d}}{K} f(pKN) \tag{5.4}$$

mit

$$f(u) = \begin{cases} u, & u \ll 1 \\ \ln(u)/u, & u \gg 1 \end{cases} \tag{5.5}$$

Der in 5.4.2 definierte Clusteringkoeffizient ist dann Gl 74

$$C(p) = \frac{3K(K-1)}{2K(2K-1) + 8pK^2 + 4p^2K^2} \tag{5.6}$$

Für die *degree*-Verteilung ergibt sich der in Abbildung 5.2 nach [6] ermittelte Verlauf.

[14] In Abschnitt 7.4.1 werden wir diese Beziehung nutzen, um die unterschiedliche Phasen der Entwicklung unser Netzwerke besser einschätzen zu können.

[15] Für Beispiele siehe Table 1 aus [6], bzw. siehe 5.2 und Table 1 aus [164].

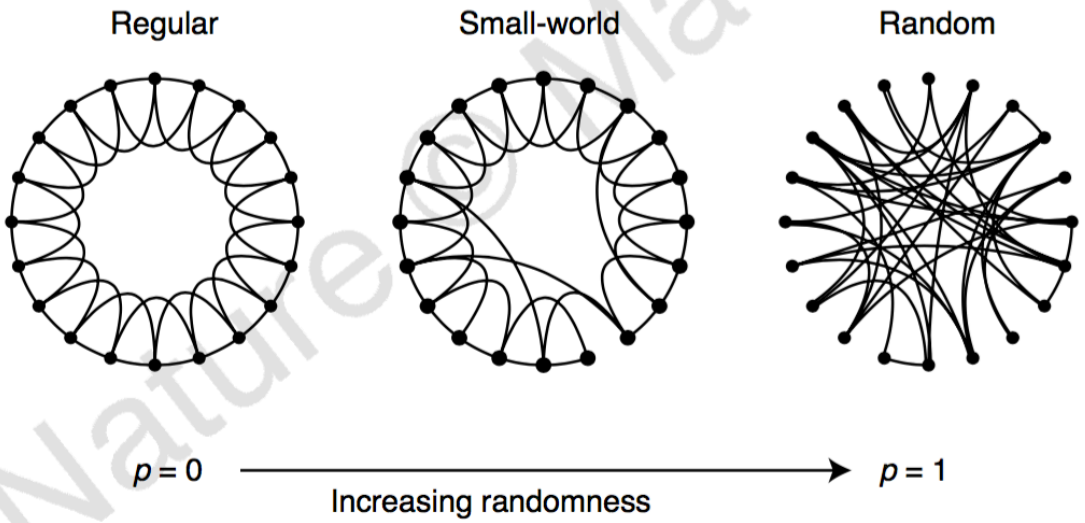


Figure 1 Random rewiring procedure for interpolating between a regular ring lattice and a random network, without altering the number of vertices or edges in the graph. We start with a ring of n vertices, each connected to its k nearest neighbours by undirected edges. (For clarity, $n = 20$ and $k = 4$ in the schematic examples shown here, but much larger n and k are used in the rest of this Letter.) We choose a vertex and the edge that connects it to its nearest neighbour in a clockwise sense. With probability p, we reconnect this edge to a vertex chosen uniformly at random over the entire ring, with duplicate edges forbidden; otherwise we leave the edge in place. We repeat this process by moving clockwise around the ring, considering each vertex in turn until one lap is completed. Next, we consider the edges that connect vertices to their second-nearest neighbours clockwise. As before, we randomly rewire each of these edges with probability p, and continue this process, circulating around the ring and proceeding outward to more distant neighbours after each lap, until each edge in the original lattice has been considered once. (As there are $nk/2$ edges in the entire graph, the rewiring process stops after $k/2$ laps.) Three realizations of this process are shown, for different values of p. For $p = 0$, the original ring is unchanged; as p increases, the graph becomes increasingly disordered until for $p = 1$, all edges are rewired randomly. One of our main results is that for intermediate values of p, the graph is a small-world network: highly clustered like a regular graph, yet with small characteristic path length, like a random graph. (See Fig. 2.)

Abbildung 5.1: Beschreibung der Bildung und Struktur von *small worlds* von Watts und Strogatz Figure 1 in [258].

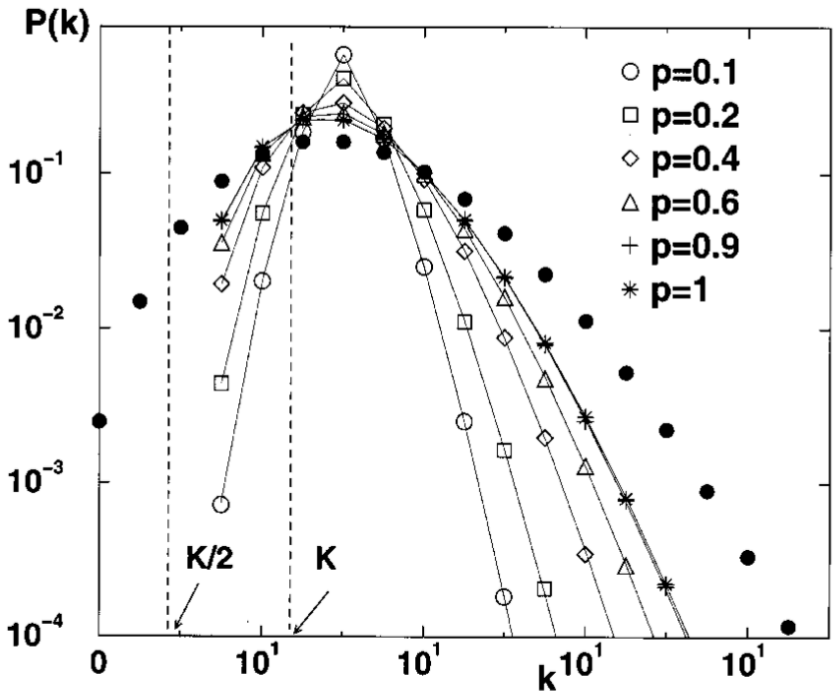


FIG. 19. Degree distribution of the Watts-Strogatz model for $K=3$ and various p. We can see that only $k \geqslant K/2$ values are present, and the mean degree is $\langle k \rangle = K$. The symbols are obtained from numerical simulations of the Watts-Strogatz model with $N=1000$, and the lines correspond to Eq. (77). As a comparison, the degree distribution of a random graph with the same parameters is plotted with filled symbols. After Barrat and Weigt (2000).

Abbildung 5.2: Fig. 19 in [6]

$$P(k) = \sum_{n=0} \min(k - K/2, K/2) C^n_{K/2} (1-p)^n p K/2 - n \frac{(pK/2)^{k-K/2-n}}{(k-k/2-n)!} e^{-pK/2} \tag{5.7}$$

für $k \geq K/2$.

In unseren Fallbeispielen werden wir nicht gezielt auf Small-World-Phänomene eingehen. Dies würde einerseits den Rahmen dieser Arbeit sprengen, andererseits ist dies der Tatsache geschuldet, dass wir, anders als in anderen historischen Untersuchungen, das Phänomen bisher nicht klar in unseren Netzwerken bestimmen können, obwohl die prinzipiellen Voraussetzungen für diese Phänomene in unseren Kontexten durchaus gegeben sind und sich diese aus der soziologischen Betrachtung unserer Netzwerke ausbilden sollten.

5.6 Exponential Random Graph Models (ERGM)

In der sozialen Netzwerkanalyse ist die Untersuchung spezieller Modelle in den letzten Jahren äußerst populär geworden. Dieses sind die sogenannten *Exponential Random Graph Models* (*ERGM*). Unter anderem existiert ein einfach zu benutzendes Paket des Open Source Statistikprogramms R.[16] ERGM erlaubt es, Hypothesen über Einflussfaktoren bei der Bildung von Beziehungen zu testen. Die Grundlagen von ERGM werden in [108] ausführlich beschrieben. Für unsere Zwecke sind jedoch einige Beobachtungen wichtig. Wir werden ERGM in den Fallbeispielen vereinzelt anwenden, jedoch gilt auch hier, dass wir die Ergebnisse wieder als heuristisches Instrument verstehen, das uns auf mögliche Probleme bei der Bildung unserer Modelle und vor allem auch bei der Erfassung und Sammlung von Daten aufmerksam machen kann. Modelle zu formulieren, die angewandt auf die historischen Daten nicht degeneriert sind, ist nicht trivial und wir erhalten häufig keine im statistischen Sinne verläßlichen Aussagen. Die Gründe sind in der Regel fundamentale Änderungen der Dynamik während der Bildung der Netzwerke, bedingt durch die sich verändernden Randbedingungen, in denen sich das Netzwerk entwickelt. Insofern ist es aus der historischen Perspektive interessant, durch die Änderung äußerer Parameter, z. B. der Länge der Zeitsamples, zu untersuchen, für welche Entwicklungen es gelingt, nicht degenerierte Modelle zu erzeugen. Ziel ist es, dadurch Hinweise auf diese fundamentalen Änderungen der Dynamik zu bekommen. Bei diesen Experimenten zeigt sich zugleich der Vorteil, den eine unterliegende semantische Modellierung für diese Form der Untersuchungen hat. Wir sind in der Lage, durch wohldefinierte semantische Abfragen unserer Wissensbasis Netzwerke zu erzeugen, die historisch interpretiert werden können. Somit wird durch das Wechselspiel von Simulation und semantischer Modellierung die Voraussetzung geschaffen, historische Hypothesen zu testen.[17]

[16] Siehe Abschnitt 6.5 und speziell zum Paket *ergm* [114].

[17] Mehr zu Darstellung von Ergebnissen in R siehe 6.5

5.7 SIENA-Modelle

Eine weitere Herangehensweise an dynamische Modelle wird durch das SIENA-Modell eröffnet. In Form von RSIENA steht auch hier ein Paket für R zur Verfügung.[18] Auf der Webseite zu RSIENA finden sich alle wesentlichen Erklärungen zu den Grundlagen des Modells. Daher auch an dieser Stelle, wie schon im vorhergehenden Abschnitt, nur das in unserem Kontext Wichtige. Analog zu dem zu ERGM Gesagten gilt auch hier, dass wir häufig keine konvergierenden Modelle auf der Grundlage unserer Daten und ersten heuristischen Annahmen erhalten, so dass auch hier die Herausforderung darin besteht, die Ergebnisse einerseits zu interpretieren, andererseits Verbesserungen am Modell vorzunehmen. Die Grundidee von SIENA ist ähnlich der von ERGM, ein generatives Modell zu finden, das in Simulationen möglichst nahe an den tatsächlichen Daten liegt. Anders als in ERGM, bei dem wir jeweils nur ein Netzwerk betrachten, betrachten wir hier die Entwicklung eines Netzwerkes in mehreren Schritten (Wellen). Auch hier ist dieses Modell nur begrenzt für unsere Fälle anwendbar. Eine der Grundvoraussetzungen ist auch hier, dass es nur eine eng begrenzte Anzahl von Akteuren gibt, die das Netzwerk in jeder der Phasen verlassen oder neu hinzukommen. Für unsere Fallbeispiele gilt jedoch, dass wir eine hohe Fluktuation der Beteiligten beobachten. Aber auch hier gilt, dass wir mithilfe der Modelle Hypothesen zu generieren versuchen, die uns dabei helfen können, die historischen Prozesse, die hinter der Formation der Netzwerke stehen, besser zu verstehen. Ein Ansatz hierbei ist es, die mit anderen Methoden gewonnenen Einsichten mittels SIENA zu validieren. Wir werden später sehen, dass sich Entwicklungen in unterschiedlichen Phasen, für die wir auf Grund der historischen Analyse unterschiedliche Entwicklungen annehmen, dann modellieren lassen, wenn wir genau in dieser Phase die Modelle getrennt berechnen. Das Gesamtmodell wird dann zu einem Randwertproblem für das Zusammenfügen der einzelnen Teile. Zukünftig ist interessant, inwieweit auch der umgekehrte Weg gegangen werden und über Singularitäten im Modell auf historische Umbrüche geschlossen werden kann.

5.8 Multilevel-Netzwerke

Reale Netzwerke lassen sich in der Regel weder durch monomodale noch bipartite Graphen darstellen. Sie besitzen häufig komplexere Strukturen. Diese Netzwerke bezeichnen wir, Snijders und Lazega [138] folgend, als *Multilevel-Netzwerke*. Diese lassen sich jeweils in eine Menge von bipartiten und monomodalen Graphen zerlegen. Die charakteristischen Größen dieser Graphen sind dann jedoch in der Regel nicht unmittelbar zu interpretieren, da sie miteinander in Wechselwirkung stehen. Sind diese Graphen ungerichtet, so lassen sich immer Graphen einer Knotenart erstellen, wenn wie bei der bipartiten Projektion jeweils zwei Kanten durch eine ersetzt werden. Der Terminologie von Snijders folgend, nennen wir diese erweiterte monomodale Graphen oder auch lediglich erweiterte Graphen.

Abhängig von ihrer Interpretation und Struktur werden unterschiedliche Typen von Netzwerken unterschieden. Eine Typologie der wesentlichen Interaktionsmodelle in multimodalen Graphen hat Elisa

[18] Siehe Abschnitt 6.5 und speziell zu Siena [222] und [220].

Bellotti in [19, Tabelle 9.1] zusammengefasst. Für uns von Bedeutung (vor allem in Abschnitt 7.2) sind hierbei Netzwerke, die mittels des Multiple-Membership-Multiple-Classification-Ansatzes (MMMC)[19] und der Erweiterung des bereits geschilderten ERGM auf Multilevel-Netzwerke in Form der *Multilevel Network Analysis* (*MNA*)[20] beschrieben werden können.

5.9 Entwicklungsdynamiken und die Etablierung wissenschaftlicher Felder

Im Kontext der Netzwerkanalysen hat sich in den letzten Jahren ein Feld aufgetan, das die Etablierung wissenschaftlicher Disziplinen und Forschungsfelder anhand einer Kombination von Zitationsanalysen und sozialer Netzwerkanalyse untersucht. In unseren Fallbeispielen untersuchen wir unsere Entwicklungsmodelle auch vor dem Hintergrund dieser Entwicklungen. Das Modell, das wir hierbei explizit mit unseren Daten vergleichen, wurde von Bettencourt und Kaiser in [21, 23] entwickelt. Der dort gemachte, in erster Linie beobachtungsgetriebene Versuch einer Beschreibung der Entwicklung von Forschungsfeldern durch den Vergleich von statistischen Größen und historischen Annahmen zeigt zumindest plausible Korrelationen von Netzwerk- und Forschungsfeldentwicklung auf, die uns veranlassten, unsere Daten in analoger Weise (Abschnitt 7.6.1) zu analysieren. Auch hier gilt für uns wieder, dass wir daraus keine Erklärung von Ursachen erwarten, sondern Hinweise auf näher zu untersuchenden Abläufe gewinnen wollen.

Das Modell von Bettencourt und Kaiser setzt im Wesentlichen das Wachstum der größten Komponente und die Anzahl der Akteure miteinander in Beziehung. Für Details sei auf die oben zitierten Artikel verwiesen. Wesentlich sind hierbei die folgenden zwei Indikatoren, die wir in diesem Kapitel anwenden werden.

Der erste Effekt, den wir untersuchen, ist der von Bettencourt und Kaiser als *densification* bezeichnete Effekt. Hier betrachten sie [23, Gl. 2] ein *power law* für die Beziehung zwischen Kanten und Knoten für ein sich über die Zeit entwickelndes Forschungsfeld:

$$\text{edges} = A(\text{nodes})^{\alpha}.$$

Eine empirische Untersuchung unterschiedlicher Fallbeispiele zeigt hierbei eine gute Bestätigung für die Gültigkeit einer solchen Beziehung für unterschiedliche Forschungsfelder ab dem Zeitpunkt, an dem ein Forschungsfeld sich formiert hat. Kanten sind hier durch Ko-Autorschaft und die Knoten durch die Autoren vorgeben. Es gilt außerdem $\alpha > 1$. Je weiter dieser Faktor von 1 entfernt ist, desto stärker ausgeprägt ist das Forschungsfeld.[21]

Einen weiterer Effekt ist die Abhängigkeit des Radius der größten Komponente von der Anzahl der Knoten in der größten Komponente. Hier dient das Ausbilden eines Plateaus ab einem Punkt in

[19] Siehe [19, S. 217] bzw. dort zitiert [35].

[20] Siehe [19, S. 217] bzw. dort zitiert [257], sowie [256].

[21] Methodische Probleme mit diesem Ansatz und die Frage, inwieweit Auswahleffekte hier eventuell zyklische Schlüsse ergeben, diskutiere ich hier nicht. Für unseren heuristischen Ansatz ist dies insofern zunächst nicht wesentlich, da auch wir bisher lediglich auf der Suche nach Indizien für Forschungsfeldformation sind und die Ansätze von Bettencourt und Kaiser in dieser Weise betrachten.

der Entwicklung der größten Komponente als ein Kriterium für die Stabilisierung eines Feldes. Auch diesen Effekt untersuchen wir in unserer Fallstudie.

5.10 Kozitationen, Forschungsdynamik und Burstness

Zur Untersuchung von Forschungsdynamiken schlägt Chen in [41],[44] und [42] ältere Studien von Pearson und Price aufgreifend die Untersuchung von Kozitationen vor. Er nimmt hier das Konzept von *research front* und *intellectual base* auf, um das Aufkommen neuer Forschungsbereiche zu detektieren und insbesondere die zentralen Inhalte aufkommender neuer Themenfelder zu charakterisieren. Der wesentliche Aspekt dieses Ansatzes ist das Zusammenführen von Text- und Netzwerkanalysen. Dabei werden Cluster von Artikeln gebildet, die zu einem bestimmten Zeitpunkt bzw. in einem bestimmten Zeitabschnitt gemeinsam zitiert werden. Diese Cluster bilden die intellektuelle Basis für die *research front* in diesem Zeitabschnitt. Die Struktur der *research front* wird dann durch ein bipartites Netzwerk aus zentralen Termen und Artikeln der *intellectual base* gebildet. Die Netzwerke „contain three types of links: (a) co-occurring research front terms, (b) co-cited intellectual base articles, and (c) a research front-term citing an intellectual base article.“[41, S.365]. Der Ansatz geht davon aus, dass sich aus der zu einem bestimmten Zeitpunkt gleichzeitig zitierten Literatur Aussagen über die zu diesem Zeitpunkt vorhandene Wissensbasis und aus der zitierenden Literatur Aussagen über die aktuelle Forschungsfront treffen lassen. Chen selbst gibt in seinen Veröffentlichungen eine Reihe von Beispielen an. Heiner Fengerau demonstriert die Stärke dieser Methode erfolgreich am Netzwerk von Jaques Loeb [79]. Das von Chen vorgeschlagene und in seinem Programm *CiteSpace II* implementierte Verfahren hat zugleich die Stärke, mittels des in Abschnitt 5.10 beschriebenen Verfahrens sprechende Titel für die sich aus den Clustern des Kozitationsnetzwerkes ergebenen Gruppen von Artikeln zu generieren. In 7.7 zeigen wir, dass sich dieses Modell mit Erfolg auch auf das Feld der Forschung zur allgemeinen Relativitätstheorie übertragen lässt. In 8.8 weiten wir den Ansatz auf die Untersuchung von Kommissionen aus, um die Etablierung von neuen Forschungsfeldern zu untersuchen.

Im Kontext des einführend in 5.13 dargestellten Ansatzes des *epistemischen Netzwerkes* ist der Ansatz von Chen ein Beispiel dafür, wie die Verbindung zwischen semantischem, sozialem und semiotischem Netzwerk praktisch vollzogen werden kann.

Die Kozitationen sind hier Ausdruck eines sozialen Netzwerkes arbeitender Wissenschaftlerinnen und Wissenschaftler, deren Ergebnisse sich formal im *semiotischen Netzwerk* der zitierten Artikel niederschlagen, während die Bildung von inhaltlichen Clustern auf Grundlage der Auswertung der Netzwerke und der transportierten Inhalte durch die Textanalyse Hinweise auf das dahinter liegende *semantische Netzwerk* geben.

Die Besonderheit des Ansatzes von Chen ist hierbei, dass die Kategorisierung von aufgegriffenen Inhalten, die in einem Netzwerk diskutiert werden, mit der klassischen sozialen Netzwerkanalyse zusammengeführt wird. Dazu greift Chen das von Kleinberg [131] eingeführte Konzept der *burstness* auf. Kleinberg entwickelt hierbei sein Konzept ausgehend von der Frage, wie sich in großen Textmengen, die zeitlich geordnet eingehen, das Aufkommen besonders relevanter Themenfelder

erkennen lässt – wie also relevante Kernbegriffe vom allgemeinen Untergrundrauschen unterschieden werden können. Er wendet hierbei seine Methodik erfolgreich auf mehrere Fallbeispiele an. Eine der entscheidenden Fragen ist dabei, wie eine sinnvolle Zeitordnung der eingehenden Daten vorgenommen werden kann. Die Wahl der Zeitordnung hängt hier unmittelbar von der intendierten Interpretation der erkannten Begriffe ab. Bei wissenschaftlichen Artikeln liegt die Vermutung nahe, dass das Einreichungs- oder Erscheinungsdatum auf die Relevanz von Begriffen zum Zeitpunkt der Einreichung bzw. der Zeit der Abfassung von Dokumenten eher relevant für Aussagen über die *research front* sind, während die Rezeptionszeiten und insbesondere die Kombination von Dokumenten auf die Entwicklung gemeinsamer Begriffe der *intellectual base* hinweisen. Chen setzt hierbei auf Kozitationen als relevantes Ordnungsmerkmal für die zeitliche Ordnung der Relevanz von Begriffen, zusätzlich trägt die Stellung der Artikel im Koziationsnetzwerk neben der Bewertung der *burstness* zur Bewertung eines Begriffes bei. Die Fallstudie zur allgemeinen Relativitätstheorie (ART) zeigt, dass dieser Ansatz auch für die dort betrachteten Zitationsnetzwerke zu einer bemerkenswerten Übereinstimmung der computergestützten Analysen mit den Hypothesen der Wissenschaftler führt, die sich in Detailstudien mit der Entwicklung der ART beschäftigt haben.

In dem von uns später betrachteten Beispiel der Relevanz von Institutionen bzw. handelnden Gruppen innerhalb einer Institution stehen wir vor einer ähnlichen Herausforderung, wenn wir neben Macht- und Einflussbeziehungen im sozialen Netzwerk, also der Frage, wer Entscheidungen beeinflusst hat, uns den Themenfeldern nähern wollen, die durch diese sozialen Netzwerke befördert wurden, und untersuchen wollen, auf welcher inhaltlichen Grundlage diese basieren. Neben der auf den historischen Quellen basierenden Konstruktion der Netzwerke stellt sich hierbei das Problem, welches die Dokumente sind, aus deren Auswertung auf Inhalte geschlossen werden kann.

Stellt sich auch bei der Auswertung von Zitationsnetzwerken die Frage nach der prinzipiellen Verfügbarkeit von Texten, zum Beispiel bei der Industrie- oder Militärforschung, und einer dadurch verursachten Verzerrung und möglichen zeitlichen Verzögerung beim Erkennen von Themenfeldern sowie nach dem Einfluss von nicht verschriftlichten Beiträgen, wie bei Vorträgen auf Konferenzen oder informellen Workshops, so gilt dies natürlich umso mehr für die Analyse von Netzwerken, die neben in Protokollen formell festgehaltenen formalen Diskussionen und Entscheidungen immer auch einen informellen Anteil haben müssen.

Wir müssen daher, insbesondere wenn es um wissenschaftliche Inhalte geht, von der Annahme ausgehen, dass relevante Einflussfaktoren in den Dokumenten erwähnt werden – die Analyse der offiziellen Dokumente über Themenfelder wird dann einen besseren Eindruck liefern als über die Auswahl von konkreten Personen. Auch diese Aussage ist selbst natürlich nur mit Einschränkungen aufrechtzuerhalten, insbesondere wenn ein Kriterium bei der Auswahl von Personen für bestimmte Positionen, persönliche Exzellenz und nicht nur das Thema selbst ausschlaggebend sein sollte.

Dennoch sehen wir in der im Fallbeispiel der MPG vorgenommenen Ausweitung des Ansatzes von Chen einen gerechtfertigten Ansatz zur weiteren Analyse der Dokumente.[22]

[22]Eine Übertragung der Methoden von Chen auf unsere Fragestellung liegt auch auf der allgemeinen Ebene unserer Theorie der epistemischen Netzwerke nahe. Hier geht es um eine Ausweitung der in den Arbeiten von Chen lediglich auf der Grundlage von Veröffentlichungen rekonstruierten Wissensbasis, um die von uns angenommenen Anteile des semiotischen Netzwerkes, die sich aus anderen Formen der kodifizierten Übermittlung von Wissen, wie Experimentalsystemen,

Die theoretischen und mathematischen Grundlagen zu *burstness* finden sich in [131]. Die Implementation, die den Auswertungen in den Beispielen zugrundeliegt, ist neben dem von Chen für die Zitationsanalyse zur Verfügung gestellten Programm *CiteSpace II* für die Kommissionen eine Adaption des von Erik Peirson maßgeblich entwickelten Pythonpaketes *Tethne* [103].

Kleinbergs Ansatz beruht auf der Grundidee, in einem kontinuierlichen Strom von eingehenden Daten Momente besonderer Aktivität zu bestimmen. Ausgangspunkt ist die Theorie *autonomer Automaten*. Von einem Ausgangszustand, d. h. einem Satz von eingegangen Nachrichten mit einem vorgegebenen Inhalt, wird dann der wahrscheinlichste Zustand ermittelt, der sich unter bestimmten Annahmen aus diesem Zustand entwickelt. Dieser Zielzustand dient dann dazu, die Relevanz des Ausgangszustandes für die Ausbreitung einer Information zu bestimmen. Die Menge dieser Zustände bestimmt den *Informationsraum*. Dazu entwickelt Kleinberg auf der Grundlage von *hidden markov models* für *endliche und unendliche Automaten* eine Kostenfunktion für die Übergänge zwischen den unterschiedlichen Zuständen unseres Informationsraumes. Hierbei wird grundsätzlich angenommen, dass Übergänge von höheren Zuständen im Sinne eines zu niedrigeren Zuständen keine Kosten erzeugen. Ein Zustand ist hierbei *höher* als ein anderer, wenn er mehr relevante Daten in Bezug auf den vorgegebenen Inhalt enthält, oder anders ausgedrückt eine höhere Emissionsrate von relevanten Nachrichten besitzt.[23] Kern hierbei ist die rekursive Formel zu Bestimmung der minimalen Kostenfunktion [131, S.382]:

$$C_j(t) = -l \ln f_j(x_t) + \min_l(C_l(t-1) + \tau(l,j))$$

mit $C_0(0) = 0$ und $C_j(0) \to \infty$ für $j > 0$.

$\tau(l,j)$ beschreibt hierbei die Kosten für einen Übergang des diskreten Zustandes l in den Zustand j [131, S.380].

$$\tau(l,j) = \begin{cases} (j-l) * \gamma * \ln(n) & \text{Wenn } j > l \\ 0 & \text{sonst} \end{cases}$$

n beschreibt hierbei die Anzahl übermittelten Botschaften z.B. die im Zeitraum T veröffentlichten Dokumente und γ einen freien Parameter, der die Kosten für einen einzelnen Übergang bestimmt. Im Allgemeinen wird dieser auf 1 gesetzt [131, S.381]. Größere Werte erschweren das Auftreten eines *burst*. Es zeigt sich, dass die Anzahl der maximal möglichen Zustände für j auch in einem unendlichen Automaten im Normalfall auf eine kleine Zahl ($k \approx 25$) reduziert werden kann, ohne dass sich das Ergebnis wesentlich ändert [131, S.382].

Ein Maß für die Relevanz bzw. Stärke eines *burst* ist dann das maximale Intervall, in dem sich der optimale Zustand in einem Zustand größer als ein vorgegebener Wert i befindet.

Hierarchien von Zuständen lassen sich unmittelbar daraus ableiten, indem wir Zustände mit $i+1, i+2, \ldots$ betrachten. Da die Menge der Zustände mit $i+1$ in denen der Zustände mit i enthalten sind, ergibt sich so eine Hierarchie von Zuständen immer höherer Stärke.

Artefakten oder formalen Sprachen ergeben. Dieses entspricht einer verallgemeinerten Wissensbasis im Sinne von Chen. Ich komme am Ende dieser Arbeit in 8.8 noch einmal darauf zurück.

[23] Hierbei ist zunächst einmal nicht erheblich, wie diese Relevanz bestimmt wird.

5.11 Epistemische Systeme

In mehreren Arbeiten [97, 100, 99] zur Entdeckungsgeschichte des Harnstoffkreislaufes durch Hans Krebs und Kurt Hanseleit beschreiben Gerd Graßhoff und Michael May das Konzept *epistemischer Systeme*. Diese sind die Grundlage für das von ihnen entwickelte Computerprogramm[24] zur Rekonstruktion des Entdeckungsprozesses. Die zitierten Arbeiten geben eine detaillierte Einführung in die von Gerd Graßhoff verfolgte Methodik. Im Kontext dieser Arbeit dient die kurze Darstellung dazu, noch einmal zu verdeutlichen, welche Ziele wir mit dem netzwerk- bzw. modellierungstheoretischen Ansatz verfolgen, der in dieser Arbeit dargestellt wird. Wir sehen hier, wie ein von konkreten Aktivitäten abstrahierendes Handlungsmodell sich unmittelbar in eine graphentheoretische Darstellung einbetten läßt. Mit den folgenden Bausteinen erhalten wir ein Handwerkszeug, dass einen ersten formalen Schritt in Richtung einer Beschreibung des zu Beginn geschilderten komplexe Systems von Netzwerken darstellt.

5.11.1 Grundbausteine eines epistemischen Systems

Ein *epistemisches System* besteht aus den folgenden Komponenten:

- *Epistemische Ziele* (*epistemic goals*): Zielvorgaben, die sich der Wissenschaftler setzt und an denen er den Erfolg seines Vorhabens misst.
- *Propositionale Einstellungen* (*propositional attitudes*): Annahmen, Überzeugungen und Vorwissen
- *Heuristik* (*heuristics*): Handlungsregeln, nach denen neue propositionale Einstellungen erzeugt und bestehende modifiziert werden, sowie Regeln zum Test von Ergebnissen theoretischer Überlegungen, wie Experimente, Auswertungen von Beobachtungen und Gedankenexperimente.
- *Epistemische Handlungen* (*epistemic actions*): Konkretisierungen von vorgeschlagenen Handlungen, die sich aus der Heuristik ergeben.

5.11.2 Kausale Graphen

Die Dynamik des Entdeckungsprozesses lässt sich als *kausaler Graph* darstellen.

Struktur der Graphen

Ein kausaler Graph ist hierbei ein gerichteter Graph, der aus einem Satz von Grundbausteinen aufgebaut ist. Diese sind:

[24] Das Programm steht leider zum Zeitpunkt des schriftlichen Verfassens dieser Arbeit nicht mehr zur Verfügung `http://www.philoscience.unibe.ch/docuserver/echo/projekte/index/modeller.html`.

Kausale Ketten	B folgt zwangsläufig aus A.	$A \rightarrow B$
Mehrfacheffekte	B und C folgen zwangsläufig aus A.	$A \rightarrow (B \wedge C)$
Mehrfache Ursachen	Sowohl A als auch B erzeugen C bzw. umgekehrt C folgt sowohl aus A als auch aus B.	$A \vee B \rightarrow C$
Komplexe Ursachen	A und B müssen gleichzeitig erfüllt sein, um C zu erzeugen.	$A \wedge B \rightarrow C$
Verhindernder Faktor	A verhindert B	$A \rightarrow \neg B$
Kausale zyklische Prozesse	Aus A folgt B folgt A.	

Minimalität

Ein zentrales Organisationsprinzip zur Selektion von Kanten des Graphen ist die Idee der Relevanz. In komplexe Ursachen werden ausschließlich Wirkungen einbezogen, die für die Wirkung B notwendig sind. Formal ausgedrückt gilt also für jede Teilursache A_i einer komplexen Ursache $\mathfrak{A}$: $\neg A_i \rightarrow \neg B$ oder andersherum ausgedrückt impliziert die Wirkung B das Vorhandensein jeder Teilursache A_i von $\mathfrak{A}$, d.h $B \rightarrow A_i$. Zusätzlich gilt: Keine einzelne Ursache A_i und keine echte Teilmenge einer komplexen Ursache $\mathfrak{A}$ ist alleine hinreichend für die Wirkung B, d.h. für kein einzelnes A_i in $\mathfrak{A}$ gilt $A_i \rightarrow B$. Eine Theorie, die diese Bedingung erfüllt, heißt *minimale Theorie*.

Handlungsprinzip

Das *Handlungsprinzip (principle of action)* bestimmt die Beziehung zwischen Ziel, Vorwissen, Fähigkeiten, Intention und konkreter Handlung. Gegeben sei eine Person A, eine (komplexe) Handlung H, sowie ein Ziel A, dann wird das *Handlungsprinzip* durch die folgende kausale Aussage ausgedrückt:

1. A hat das Ziel G,
2. A glaubt, dass H unter den gegebenen Umständen, ein Weg ist um G zu erreichen,
3. Es gibt nach Meinung von A keinen anderen Weg höherer Priorität, um G zu erreichen,
4. Es gibt kein anderes Ziel G' (mit höherer Priorität), dass ihn davon abhält G zu realisieren,
5. A weiß wie H auszuführen ist,
6. A besitzt die Fähigkeit H auszuführen,

dann konkretisiert A die Handlung H.

Konkretisieren meint in diesem Sinne: Person A beginnt mit der konkreten Umsetzung der Handlung.

5.11.3 Epistemischer Handlungsraum

Kernpunkt für die von Graßhoff und May durchgeführte Umsetzung der Handlungsstrategien von Krebs und Hanseleit in ein Computermodell ist der *epistemische Handlungsraum* (*epistemic action space*).

Zur Erreichung des übergeordneten Ziels – der Lösung des gestellten Forschungsproblems – wird das Problem in einzelne Unterprobleme zerlegt, die dann in einer Kombination von heuristischen Suchverfahren und Mechanismen zum Test und zur Revision der gewählten Verfahren angegangen werden. Bei jedem Schritt ist nur ein Ausschnitt des gesamten Handlungsraumes sichtbar. Jeder einzelne Schritt wird jeweils so lange expandiert, bis er in einem Knoten endet, der in einfacher Weise evaluiert werden kann, z. B. die Planung und Ausführung eines Experimentes. Die vorangegangenen Schritte werden auf einem Stapel abgelegt. Bei jedem Entscheidungsschritt wird nur der lokale Raum betrachtet und in diesem Raum nach Prioritäten (*preference rules*) entschieden. In diesem Punkt unterscheidet sich das Verfahren vom klassischen Best-First-Search Verfahren, bei dem für die Bewertung eines Pfades jeweils der gesamte Pfad betrachtet wird.

Struktur des lokalen Handlungsraumes

Der lokale Handlungsraum spaltet sich in primitive Handlungen und komplexe Handlungen auf. Primitive Handlungen sind Handlungen, die nicht weiter aufgeteilt, sondern direkt ausgeführt werden und „mit erfolgreich“, „nicht erfolgreich”’ oder „noch nicht ausführbar“, weil erst noch ein anderer Schritt ausgeführt werden muss, evaluiert werden können. Komplexe Handlungen sind Handlungen, die in weitere Einzelschritte zerlegt werden müssen. Die Unterscheidung zwischen komplexen und primitiven Handlungen liegt nicht unveränderbar fest, sondern hängt vom Detailgrad des Modells ab.

Komplexe Handlungen teilen sich zwischen konkreten Handlungen im Rahmen des Experimentalsystems und Schritten im epistemischen System auf. Erstere beschreiben hierbei konkrete Handlungen, wie die Ausführung eines Experimentes oder den Vergleich von theoretischen Ergebnissen mit vorhandenen Messdaten. Letztere beschreiben notwendige Modifikationen der Theorie, diese lassen sich nach Graßhoff und May auf drei Typen reduzieren. Im Sinne der eingangs in Abschnitt 5.13 geschilderten Theorie des *epistemischen Netzwerkes* entsprechen die durch das *Experimentalsystem* fest vorgegeben Schritte Handlungen im *semiotischen Netzwerk*, während Theoriebildung einen Prozess innerhalb des semantischen Netzes abbildet.

- *Expansion*, d.h. der Theorie wird eine weitere Restriktion oder Annahme hinzugefügt
- *Kontraktion*, d.h. eine Restriktion oder Annahme wird aus der Theorie entfernt
- *Revision*, d.h die Kontraktion mit anschließender Expansion.

Implementierungsregeln

Zusammenfassend muss also eine Implementation eines epistemischen Systems die Bausteine des epistemischen Systems, wie in 5.11.1 beschrieben,formalisieren. Konkret bedeutet dies, es wird eine

- Repräsentation der Heuristik benötigt, insbesondere des lokalen Handlungsraumes, sowie
- eine Formalisierung der Präferenzregeln analog den Methoden des Best-First-Search-Algorithmus mit den oben (in 5.11.3) beschriebenen Modifikationen.
- Vorhandenes Basiswissen muss in Form einer Wissensdatenbank zur Verfügung stehen.
- Ergebnisse von Experimenten müssen bereit gestellt werden.

5.12 Historische Netzwerkforschung – eine Übersicht

Historische Netzwerkforschung ist in den letzten Jahren zu einem schnell wachsenden Feld geworden. Die Webseite „Historical Network Research“ [67] gibt einen umfassenden Eindruck über laufende Forschungen und spiegelt die Dynamik des Feldes wider.

Netzwerke werden und wurden in der historischen Forschung in sehr unterschiedlichen Kontexten und Anwendungsfällen benutzt. Während es eine lange Tradition gibt, Netzwerke und Graphen als qualitatives Instrument und zur Visualisierung und Veranschaulichung zu nutzen, ist die quantitative Analyse ein neues Phänomen, das in den 1990er Jahren eingesetzt hat. In die Geschichtswissenschaften hat die Netzwerkforschung zunächst in Form der Übertragung von sozialwissenschaftlichen Methoden auf historische Fragestellungen Einzug gehalten. Wie stets bei der Anwendung quantitativer Methoden auf die historische Forschung gilt es hier, die Validität der Daten im Auge zu behalten.

> „Bereits in den 1970-er Jahren wiesen Historiker darauf hin, dass deren Output vielmehr sorgfältig interpretiert werden müsse: „[…]hardware and software are only tools not end in themselves. In historical research the intellectual question must always be dominiant“. Bei aller Nützlichkeit sozialwissenschaftlicher Methoden zähle letzten Ende immer noch die historische Synthese die „außerhalb der Reichweite der Sozialwissenschaften“ [25] liege, weder die Handlungen historischer AkteurInnen noch komplexe historische Ereignisse oder Ideen seien letztendlich quantifizierbar.“[26]

Typische Vertreter für die Anwendung qualitativer Methoden aus der Netzwerkforschung sind sicher die Darstellung von Herrschafts- und Handelsbeziehungen, Darstellungen von Vertragssituationen

[25] Zitat hier übernommen, Fogel [83, s. 324] führt weiter aus: „The task which historians set for themselves cannot be achieved through social science alone. Because historians aspire to comprehend the totality of human behavior, their concerns transcend the subject matter of the social sciences and enter moral and aesthetic realms. Even with respect to those issues which fall within the scope of social science, historians frequently demand more than social science can deliver. This is certainly the case when historians attempt to combine all the elements of human behavior that concern social scientists- economic, social, political, psychological, and cultural-into a „seamless web.“

[26] Zitat aus [130, S.37ff]. Kerschbauer und Düring beziehen sich hierbei auf [134, S.173]. Das vollständige Zitat lautet hier: „The logic of theory application in history is complex. Theories can be made fruitful for historical work in different ways. In some examples of social-scientific history, the source data are completely subordinated to the theories and the theory-derived hypotheses. In such cases, the general rules of data analysis common to the social sciences are applied. I think this practice may be legitimate, but it is not without problems and certainly it is not the only way. Most historians prefer more flexible ways of combining theory and sources. Frequently theories become just the backbone of an argument that itself contains non theoretical, descriptive, and narrative dimensions as well.“

und natürlich insbesondere Stammbäume in der genealogischen Forschung. Im Hinblick auf die Anwendung quantitativer Methoden ist die häufig zitierte Arbeit von John Gadget und Christopher Ansel von 1993 über das Herrschaftssystem und den Aufstieg der Medici Familie [181] ein methodisch aufschlussreiches Beispiel, da hier neben einfachen Charakteristiken, wie den Zentralitätsmaßen, insbesondere auch darüber hinausgehende Verfahren der Klassifikation von Knoten innerhalb von Netzwerken angewandt und historisch interpretiert werden [68, Kap. 2.4]. Methodischer Schwerpunkt unseres Ansatzes ist es, zwei Ansätze, die in den Digital Humanities bislang weitestgehend getrennt behandelt werden, zusammenzuführen. Dieses sind auf der einen Seite Fragen der Datenmodellierung auf der Grundlage von Ontologien.[27] Auf der anderen Seiten stehen Ansätze der historischen Netzwerkforschung, die sich aus der Analyse sozialer Netzwerke und der Untersuchung von Diffusionsnetzwerken im Zusammenhang der Innovationsforschung heraus entwickelt haben. Für die Anwendung quantitativer Methoden aus der Netzwerkforschung ist eine Grundbedingung, Netzwerke zu erstellen, die auf der gleichen epistemischen Ebene liegen. Es gilt also zu vermeiden, „Äpfel“ mit „Birnen“ zu vergleichen.

Die Zusammenführung von Modellierung und Netzwerkforschung ist hier ein Ansatz dafür, die Struktur von Netzwerken nachvollziehbar semantisch zu beschreiben, so dass die Interpretation quantitativer Ergebnisse nachvollziehbar wird und Charakteristiken, die sich aus der mathematischen Graphentheorie für Netzwerke ergeben, als historisch relevante Indikatoren interpretiert werden können. Diese Indikatoren, auch das soll hier noch einmal betont werden, sind in diesem Kontext ein Instrument zur Analyse von Quellen, die anders nur unvollständig erschlossen werden können. Die Instrumentarien und Quellen müssen genauso wie andere Quellen einer historisch-kritischen Betrachtung unterzogen werden, bevor daraus Aussagen über historische Sachverhalte abgeleitet werden können. Zurecht weist Claire Lemercier in ihren Arbeiten insbesondere auf die Gefahren und Risiken der Anwendung der historischen Netzwerkforschung hin, zeigt jedoch gleichzeitig den Weg zur kritischen Anwendung dieser Methoden auf.

> Es sollte bereits klar geworden sein, dass ich nicht denke, es sei unmöglich, formale Netzwerkmethoden auf historische Fälle anzuwenden, weil sie zu intrinsisch mit Theorien über die heutige Gesellschaft verbunden seien. Sie basieren beispielsweise nicht auf Annahmen, dass irgendein Individuum frei sei, seine Freunde unabhängig von sozialen Schranken zu wählen, im Gegenteil sie bieten robuste Werkzeuge um Homogamie (die Tendenz, sich mit Personen, die einem ähnlich sind, zu verbinden) oder Endogamie (die Tendenz sich mit Personen, mit denen man schon zuvor verbunden war, zu verbinden, zum Beispiel *matrimonial relinking*, die Wiederverbindung durch eine weitere Heirat zwischen zwei Familien) zu untersuchen. […] Solange wir nicht unsere professionellen Tugenden vergessen, präzise zu definieren, wonach wir suchen und die Perspektive historischer Akteure selbst soweit als möglich in Betracht zu ziehen, können wir formale Methoden nutzen, nicht nur um Beziehungen zwischen Personen, sondern auch zwischen Organisationen,

[27] Dahinter steht das Ziel, sich über die Beschreibung von Handlungs- und Wissensorganisationsmodellen mit Hilfe von Ontologien dem allgemeineren Konzept der mentalen Modelle und Frames zumindest anzunähern und diese damit einer Überprüfung mittels computergestützter Methoden zu eröffnen.

Orten und sogar Wörtern zu untersuchen. [140, S. 23ff]

5.12.1 Qualitative und quantitative Ansätze der Netzwerkforschung

Im Kontext der historischen Sozialforschung und Sozialgeschichte wird zumindest im deutschen Kontext häufig auf die sehr frühen Arbeiten von Wolfgang Reinhard „Freunde und Kreaturen“ aus dem Jahre 1979 verwiesen [198]. Der Fokus hier liegt auf den Verflechtungen von Sozialorganisation, den Veränderungen von Einflussbereichen und den sich daraus ergebenen Veränderungen der Handlungen einer Organisation, in seinem Fall der römischen Kurie.[28]

Jüngstes Beispiel für die konsequente Anwendung computergestützter Methoden, sowohl aus der Netzwerktheorie als auch Ansätze von Peter Turchin [243] aufgreifend, stellen die Arbeiten von Johannes Preiser-Kapeller dar. Genannt seien hier seine Arbeit zu den komplexen Strukturen maritimer Netzwerke von der Antike bis in die Frühe Neuzeit [190, 191, 189] sowie die ausführliche Studie zu den Herrschaftsstrukturen in Byzanz.[187, 188] Turchins Ansatz der *historical dynamics* wird hierbei kritisch aufgenommen und weiterentwickelt. Netzwerkforschung dient Preiser-Kapeller als Brückenschlag zwischen einem quantitativen Ansatz, der komplexe historische Systeme auf der Grundlage von mathematischen Größen beschreibt, und der Netzwerkforschung, die es ermöglicht sowohl Aussagen über das Gesamtsystem als auch über das Verhalten von Einzelnen zu machen. Makro- und Mikrostrukturen lassen sich so verbinden.

Unsere Vorgehensweise im folgenden ist es, auf der Grundlage von historischen Fakten, Indizien und Annahmen schrittweise ein Beziehungsnetzwerk zu konstruieren, dessen Dynamik mit netzwerktheoretischen Methoden beschrieben werden kann. Die Arbeiten von Preiser-Kappeller zur Gesellschaftsdynamik des mittelalterlichen Byzanz [188, 187] sind ein instruktives Beispiel für die Anwendung dieser Methodik, die neben den Erfolgen auch die Probleme im Umgang mit historischen Daten, ihrer Berechnung und den sich daraus ergebenden Interpretationen nachvollziehbar darstellen.

So wird dort das bei der mathematischen Modellierung auftretende Problem der in der Regel vorhandenen freien Parameter ausführlich diskutiert. Es werden unterschiedliche Ergebnisse der Modellberechnungen bei unterschiedlicher Parameterwahl diskutiert und es wird gezeigt, wie sich trotz der Ungenauigkeiten Ergebnisse erzielen lassen, die der historisch-kritischen Methode genügen.[29]

5.12.2 Ungenauigkeiten historischer Quellen

Immer wieder wird gegen den Einsatz quantitativer Methoden das Argument vorgebracht, dass fehlende und ungenaue Daten prinzipiell einen Einsatz solcher Methoden ausschließen, bzw. ihre Fehlerquoten zu hoch sind, um sinnvoll Aussagen zu treffen:

[28] Eine ausführliche Würdigung der Bedeutung der Arbeiten von Wolfgang Reinhard im Kontext der historischen Netzwerkforschung findet sich in Kap. 2.3 von [68].

[29] Insbesondere zeigt sich in diesen Arbeiten, wie das von Jürgen Kocka bereits 1984 in [134]) eingeforderte Verhältnis von Theorie, Quantifizierung und kritischer Reflexion in praktische computer-gestützte historische Forschung Eingang finden kann, wobei die computer-gestützte Methode als heuristisches Instrument und neue Quellengattung verstanden wird.

> Dies [Missing-Data] ist nicht notwendig ein Hinderungsgrund für die Anwendung netzwerkanalytischer Methoden. Die ForscherInnen sind dabei aber angehalten, die jeweilige Quellenlage, auf der die Untersuchung beruht, deutlich zu kommunizieren und auch explizit darauf hinzuweisen, dass ihre Analyseergebnisse nur vorsichtig zu interpretieren sind. Entgegen landläufiger Meinung haben auch die Sozialwissenschaften mindestens ebenso große Probleme mit „missing data“ und „no response“ bei den Datenerhebungen ihrer zumeist gegenwartsbezogenen Studien. Hierbei sind HistorikerInnen mitunter sogar im Vorteil, da sie sich bereits im Rahmen der „historischen Methode“ intensiv mit Fragen der Widersprüchlichkeit und Lückenhaftigkeit auseinandergesetzt haben.[30]

Es gilt jedoch auch, dass historische Netzwerkforschung nur zu rechtfertigen ist, wenn die Quellen, auf denen diese Analysen beruhen, weitestgehend öffentlich gemacht werden, so dass die gemachten Annahmen nachvollziehbar und nachprüfbar sind. Datenpublikation und Ergebnispublikation sind in diesem Sinne die zwei Seiten einer Medaille. Die von Bixler und Reupke gemachten Einwände teile ich daher ausdrücklich nicht: „Bedenkt man den teils erheblichen Aufwand bei der Erhebung, ist es wohl im Interesse der Forschung, aber nicht immer auch im Interesse der Forscherin/des Forschers, ihre/seine Daten direkt zu veröffentlichen. Es ist zudem nicht sichergestellt, dass die Daten bei etwaigen Sekundäranalysen auch in dem Sinne verstanden werden, in dem sie erhoben wurden. Allzu leicht können die Daten aus dem Zusammenhang gerissen und fehlinterpretiert werden. Bei historischen Daten scheint dieses Risiko umso höher, da ihr Hintergrund meist ohnehin nicht hundertprozentig sicher zu erschließen ist. Es soll hier nicht pauschal gegen die Veröffentlichung von Datensätzen plädiert werden, im Gegenteil kann sie der Forschung große Chancen bieten - letztendlich muss jedoch das Für und Wider im Einzelfall abgewogen werden.“[31] Mir scheint dies jedoch ein hoch problematisches Argument und eher auf die generelle Problemlage der Dokumentation von Forschungsergebnissen durch computergestützte Methoden hinzuweisen. Wenn auf Grundlage von computergestützten Methoden Schlüsse gezogen und im Rahmen einer wissenschaftlichen historischen Arbeit veröffentlicht werden, sind diese nur dann akzeptabel, wenn sowohl die Daten als auch die Methoden publiziert werden. Das Argument von Bixler und Reupke greift an dieser Stelle zu kurz. Das Risiko liegt nicht in der Publikation der Daten selbst, sondern in erster Linie in der Problematik der Vermittlung und Sicherung der genutzten Methoden. Hier liegt das zentrale Desiderat im Hinblick auf die Anwendung quantitativer Methoden in der historischen Forschung. Bisher sind sowohl die Kompetenzen der Deutung dieser Methoden als auch die Möglichkeit der nachhaltigen Publikation dieser Methoden häufig nicht gegeben. Dies führen sie selbst als zentrale Aufgabe innerhalb der Netzwerkforschung auf:[32]

> „Relationale Datenbanksysteme ermöglichen eine anpassbare Verarbeitung und Dokumentation von Quelleninhalten und Arbeitsprozess, jedoch ist hier der Zeitaufwand in Aufrechnung zum Ergebnis zu bringen. Die nach einer Quellenkritik notwendige Standardisierung der Quelleninhalte ermöglicht geschärfte und valide Ergebnisse.

[30] Stark: Netzwerkberechnungen, S 157
[31] Bixler/Reupke: Von Quellen zu Netzwerken, S166
[32] Bixler/Reupke: Von Quellen zu Netzwerken, S122

> Die Erhebung relationaler Daten aus historischen Quellen ist als Prozess so individuell wie ein Quellenkorpus selbst. Beim derzeitigen Stand der methodischen Entwicklung müssen netzwerkanalytisch arbeitende HistorikerInnen ihren Weg noch explorativ und diskursiv beschreiten.“

5.12.3 Bibliometrie und Wissensdynamik

Und schließlich sei noch ein weiterer Aspekt der historischen Netzwerkforschung genannt, auf den später noch im Fallbeispiel zu den Netzwerken der Relativitätstheorie[33] zurückgegriffen wird: die Anwendung von bibliometrischen Methoden zur Erforschung der Wissenschaftsdynamik.

Die sehr umfangreiche Studie von Heiner Fengerau aus dem Jahre 2010 [79] und seine darauf folgenden Arbeiten[34] liefern hierfür ein Beispiel. Intensiv werden hier die von Chaomei Chen entwickelten Methoden eingesetzt.[35] Insbesondere stellt der von Chen dargestellte Ansatz eine Möglichkeit dar, ausgehend von Personen- sowie Zitations- und Kozitationsbeziehungen Netzwerke zu konstruieren, in denen die Knoten Konstituenten eines Wissenssystems darstellen. In 5.13 wird die Bedeutung dieser unterschiedlichen Netzwerke im Rahmen des dieser Arbeit zugrundeliegenden wissenschaftstheoretischen Ansatzes genauer dargestellt.

Hinter den Überlegungen von Chen steht die Annahme, dass Kozitationen einen besseren Eindruck über die Entwicklung von neuen Themenfeldern durch Rekombination und Erweiterung existierender Forschungsfelder geben können als Ko-Autorschaft. Die Auswertungen im Fallbeispiel zur Allgemeinen Relativitätstheorie zeigt hier eine gute Übereinstimmung der durch diese Methoden gewonnenen Ergebnisse mit denen, die sich aus der detaillierten historischen Studie ergeben.

Die Frage, ob diese Ansätze auch auf andere Formen von Einflussnahme übertragbar sind, ist der Hintergrund für eine Teilstudie innerhalb der Untersuchung der Kommissionsnetzwerke, wie sie in 8.8 geschildert werden. Diese bezeichnen wir entsprechend als Koinfluenz-Netzwerke.

5.12.4 Zwischen qualitativer und quantitativer Analyse - Visual Analytics

Mit der Verfügbarkeit schneller Computer am Arbeitsplatz und insbesondere immer leistungsfähigerer Graphikkarten - nicht zuletzt dank der Gaming-Industrie - ist die Visualisierung von Netzwerken als farbenprächtige und komplexe Abbildungen in Mode gekommen. Häufig wird der damit verbundene methodische Ansatz zutreffend als *visual analytics* charakterisiert, der darauf abzielt, Visualisierung als analytisches Tool einzusetzen [129].

Im Umfeld der historischen Forschung ist sicherlich in diesem Zusammenhang das meist genannte Beispiel *Mapping the Republic of Letters* der Stanford University [146]. Ähnliche Projekte finden sich zum Beispiel am Getty Institute zum Kunsthandel der frühen Neuzeit und Moderne.[36]

[33] Siehe Kapitel 7 insbesondere Abschnitt 7.7.

[34] Siehe insbesonders [78].

[35] Siehe [43, 41] . Zur Analyse solcher Entwicklungen stellt Chen, das frei verfübare Programm CiteSpace [42] zur Verfügung.

[36] Besonders sollen hier in diesem Kontext die Arbeiten von Maximilian Schich genannt werden, siehe [213] und die dort genannten Referenzen.

Da Bilder eine hohe suggestive Kraft haben, ist es hier wichtig zu betonen, dass bei der Interpretation dieser Visualisierung stets ein ästhetisches Element zu berücksichtigen ist, das zu falschen Interpretationen führen kann. Visualisierungen erfordern im gleichen Maße wie textuelle Darstellungen eine kritische Betrachtung. Wie Texte sind auch Visualisierungen Interpretationen von Sachverhalten, in die subjektive Entscheidungen einfließen. So ist die relative Größe und insbesondere die Nähe von Objekten in graphischen Darstellungen häufig lediglich ein Hinweis auf die tatsächliche Nähe von Objekten. Selbst wenn man von der grundsätzlich schon subjektiven oder zumindest hochgradig interpretativen Festlegung von Metriken für Abstände von Objekten wie Texten, Personen oder gar Konzepten absieht, geben Artefakte der Projektion von multidimensionalen Räumen auf die Ebene (oder einen dreidimensionalen Körper) irreführende Eindrücke von Nähe.

Netzwerkvisualisierungen werden in der Praxis häufig durch manuelles Bearbeiten so verändert, dass die damit verbundene – interpretative – Aussage des Autors deutlicher wird. Gegen diese Verfahren ist zunächst nichts einzuwenden, solange die Eingriffe transparent für den Rezipienten dargestellt und somit nachvollziehbar werden. Qualitativ ist dies nicht anders zu bewerten als die Anwendung von spezifischen Algorithmen zur Manipulation von Daten, die bestimmte Effekte, zum Beispiel durch Gewichtung, stärker hervorheben sollen, oder die Beseitigung eines angenommenen Hintergrundrauschens in Daten. Auch hier werden bei der Auswahl der Methoden und insbesondere der Parameter subjektive Entscheidungen getroffen. Die Qualität der Dokumentation der benutzten Methoden ist damit ein entscheidendes Kriterium für die Akzeptanz von Visualisierungen als wissenschaftliche Publikationsform.

In diesem Kontext stehen zum Beispiel Marian Dörks Arbeiten [147]. Ein Schwerpunkt seiner Arbeiten liegt auf Methoden, die die kritische Analyse ermöglichen und dem Anwender verdeutlichen, welche Auswirkungen Parameteränderungen haben können und welches die algorithmischen Grundlagen für die visuelle Darstellung sind.

Es kann daher in diesem Kontext nicht überbetont werden, dass auch bei der Anwendung digitaler Methoden die Regeln der Quellenkritik im Rahmen der historisch-kritischen Methoden gültig bleiben. Auch hier gilt: Die Rückbindung jeder Form der Interpretation an die Ausgangsdaten ist notwendige Voraussetzung für die Wissenschaftlichkeit dieser neuen Methoden. Die Publikation von Ergebnissen, die durch computergestützten Methoden gewonnen werden, muss daher möglichst auch die Publikation der algorithmischen Methoden und der Primärdaten umfassen.[37]

5.13 Netzwerke als Hilfsmittel zur Strukturierung von Wissenssystemen

In der historischen Forschung standen in der Vergangenheit häufig Akteursnetzwerke im Vordergrund. Historische Netzwerkanalyse ist daher vielfach lediglich eine historisierte soziale Netzwerkforschung. Die in 5.12.3 genannten Methoden der Untersuchung von Kozitationen geben einen ersten Hinweis

[37] Hierbei sind mir die damit verbundenen urheberrechtlichen Probleme durchaus bewusst. Es zeigt sich bei dieser Form der historischen Forschung noch einmal besonders deutlich, welche Bedeutung Open Access für den Einsatz neuer wissenschaftlicher Methoden in der historischen Forschung hat.

auf eine Erweiterung der Netzwerktheorie auf die insbesondere für die Wissenschaftsgeschichte zentrale Frage, wie Wissensentwicklung und Personennetzwerke mit einander in Beziehung gesetzt werden können. In der *Kozitationsanalyse* werden Publikationen zu den Knoten des Netzwerkes und nicht wie im Falle der *Zitationsanalyse* Personen. Hierbei stehen Publikationen bereits für Ereignisse, in denen Wissensbereiche zusammengefügt werden. Sie werden an dieser Stelle Bausteine einer spezifischen Substruktur des Wissenssystems, das wir im Folgenden als das *semiotische Netzwerk* bezeichnen werden. Eine strukturierte Untersuchung der Wissensentwicklung setzt voraus, die komplexen Wechselwirkungen der unterschiedlichen Netzwerke zu verstehen, die das Wissenssystem ausmachen. Die Anwendung der Methoden der mathematischen Netzwerkanalyse zu diesem Zwecke setzt voraus, dass die unterschiedlichen epistemischen Ebenen eines Netzwerkes verstanden und beschrieben werden. Wissensnetzwerke sind in diesem Rahmen komplexe multimodale Netzwerke bzw. Multilevel-Netzwerke.[38] Zum jetzigen Zeitpunkt hat dieser Ansatz noch den Status einer Arbeitshypothese. Diese Arbeit soll jedoch aufzeigen, dass sich diese als Ausgangspunkt für eine konkrete Anwendung der mathematischen Netzwerktheorie bereits eignet. In [204] wird diese Hypothese ausführlich dargestellt. An dieser Stelle werden daher nur die wesentlichen Leitideen vorgestellt, die zum Verständnis des Kontextes dieser Arbeit notwendig sind. Der wissenschaftstheoretische Ansatz, der als Grundlage für die hier gemachten historischen Untersuchungen dient, wurde bereits in Abschnitt 2.2 eingeführt. In diesem Rahmen kann Wissen als kodierte Erfahrung und Handlungspotenzial aufgefasst werden. Die Kodierung von Wissen erfolgt durch kognitive Strukturen von Akteuren sowie durch externe Repräsentationen, z. B. Sprache und Schrift. Wissen hat demnach nicht nur eine kognitive, sondern auch eine materielle Dimension. Träger dieser Wissensstrukturen sind Akteure, die Wissen erzeugen, verbreiten und modifizieren. Von wesentlicher Bedeutung hierbei ist, dass die einzelnen Akteure sich ihrer Funktion im System der Wissensausbreitung und -strukturierung nicht bewusst sein müssen. Das Netz ist mehr als die Summe seiner Teile. Treibendes Moment in der Geschichte der Wissensstrukturierung ist das Netz als Ganzes, es sind nicht seine Teile. Das heißt nicht, dass Wissensausbreitung unbewusst von den Akteuren stattfindet, sondern lediglich, dass sich in den Netzwerken zunächst verborgene Potenziale für eine Neuausrichtung aufbauen können, die erst nach längeren Phasen realisiert werden, was den beteiligten Akteuren erst dann bewusst wird. Strukturen des Handelns und der Handlungssteuerung im Prozess der Wissensorganisation lassen sich so als System von Netzwerken verstehen. Dieses System umfasst Netzwerke drei grundsätzlich unterschiedlicher Typen:

- *Soziale Netzwerke*
- *Semiotische Netzwerke/externe Repräsentation*
- *Semantische/kognitive Netzwerke*

Diese Netzwerke sind aufeinander bezogen und liefern die Rahmenbedingungen für die Ausbreitung und Strukturierung der jeweils anderen. In diesem lassen sich die Netzwerke der unterschiedlichen Ebenen getrennt im Hinblick auf ihre topologischen Eigenschaften und Entwicklungsdynamik untersuchen. Die Eigenschaften und Strukturen der jeweils anderen Netzwerktypen gehen in die Überlegung

[38] Siehe 5.8.

hier als Randbedingungen für die Dynamik der Netzwerke ein, die sich im Wesentlichen als Bedingungen für die Ausprägungen der Kanten der jeweiligen Netzwerke zeigen.

***Soziales Netzwerk*.** Für eine wissensgeschichtliche oder wissenschaftshistorische Analyse stehen dabei die an der Wissensorganisation und -produktion beteiligten Akteure und ihre entsprechenden Austauschprozesse im Vordergrund; in diesem Falle sprechen wir von sozialen Wissensnetzwerken. In diesem Sinne stellen die Knoten Agenten in einem Netzwerk da, die untereinander in Beziehung stehen und Informationen austauschen. Wie wir später sehen werden, heißt dies jedoch nicht, dass die Agenten sich dieser Prozesse bewusst sein müssen. Vielmehr stellt das in dem Netzwerk gespeicherte Wissen ein Potenzial dar, das sich erst entwickeln muss und das durch das kollektive Handeln realisiert wird.

***Semiotische Netzwerke/externe Repräsentation*.** Als Zweites führen wir das Netzwerk ein, dessen Knoten die realen Gegenstände und externen Repräsentationen des Handelns sind und dessen Kanten die Handlungen oder physischen Beziehungen sind, durch die ein Zusammenhang zwischen diesen Gegenständen oder Zeichen existiert. Wir nennen es daher der Einfachheit halber das *semiotische Netzwerk*, obwohl es nicht auf Zeichen beschränkt ist, sondern den gesamten materiellen Kontext des Handelns einschließt.

Umgekehrt sind die materiellen Netzwerke – wie etwa Schriftquellen – die in der Regel einzigen Grundlagen zur Rekonstruktion des sozialen Netzwerkes, das wiederum die Voraussetzung zur Erzeugung der materiellen Repräsentation des Netzwerkes selbst ist.[39]

***Semantisches Netzwerk*.** Auch die internen Strukturen des Wissens und des normativen Denkens lassen sich als eine solche Netzwerkstruktur darstellen, die wir als das *semantische Netzwerk* bezeichnen wollen [115].

Alle diese drei Netzwerktypen bedingen sich gegenseitig, denn Handeln ist durch seine externen Bedingungen, in unserer Sprechweise also die *semiotischen Kontexte*, bestimmt und greift verändernd in diese ein. Zugleich wird das soziale Handlungsnetz u.a. durch das den Handelnden verfügbare Wissen und durch Institutionen als Mechanismen kollektiver Handlungssteuerung bedingt und wirkt auf diese zurück. Aus dieser Perspektive kann also das semiotische Netzwerk – oder zumindest seine Knoten – als eine „Externalisierung“ des sozialen Handlungsnetzwerks verstanden werden und das semantische Netzwerk als seine „Internalisierung“, die beide steuernd auf das Netz des sozialen Handelns zurückwirken [137].

[39] Schon Wolfgang Reinhard weist hierbei auf das Risiko eines Zyklus hin. „Dennoch, schaffen wir uns nicht einen methodologischen Zirkel, wenn wir Briefe zur Hauptquelle für Existenz und Struktur eines „networks“ machen, nachdem wir sie eben erst zum Produkt dieses von uns postulierten „network“ erklärt haben?“. Er beantwortet die Frage danach direkt analog zu unserem Ansatz einer notwendigen Verbindung zwischen Einzelstudie und Netzwerkuntersuchung. „Der Einwand entfällt jedoch, weil diese ’Überreste’ durch Traditionsquellen ergänzt werden, in denen uns das Funktionieren des ’network’ expressis verbis an konkreten Einzelfällen vorgeführt wird […].“[198, S. 63]

5.14 Netzwerke und Modellierung

Die Interaktionen zwischen den unterschiedlichen Ebenen lassen sich in Form von Multilevel-Netzwerken beschreiben.[40] Hinzu kommt ein weiterer Aspekt, den wir in dieser Arbeit nicht weiter verfolgen und schon in Abschnitt 1.1.2 angesprochen hatten: Mit Hilfe von Handlungsmodellen und mentalen Modellen bekommen wir eine Struktur vorgegeben, vor deren Hintergrund wir die Netzwerke beschreiben. In dieser Arbeit zeigen wir, dass semantische Modellierung ein Gerüst für die dann netzwerktheoretisch zu untersuchenden Netzwerke vorgibt. Auf der semantischen Ebene können wir, dank der entsprechenden semantischen Beschreibung der Beziehungen, Knoten der unterschiedlichen Ebenen miteinander in Verbindung setzen. Auf Beschreibungslogiken aufbauende Modellierung gibt die Einschränkungen und Abhängigkeiten vor und erlaubt das Prüfen der Modelle auf Konsistenz.

Die zentrale Frage ist die Beschreibung der Dynamik solch komplexer Systeme. Es existieren drei unterschiedliche Dynamiken: eine bestimmt die Entwicklung der Instanzen des Systems, eine die Fortschreibung der Ontologie, wenn wir diese als zeitgebundenes Ordnungssystem verstehen, und schließlich die dritte beschreibt Prozesse, die sich in den Netzwerken abspielen.

[40] Siehe 5.8.

Teil II

Fallbeispiele und Implementation

Kapitel 6

Die Arbeitsumgebung

Dieses Kapitel bietet eine kurze zusammenfassende Übersicht über die technische Infrastruktur und die in den folgenden Fallstudien eingesetzte Infrastruktur. In den folgenden Kapiteln wird jeweils auf dieses Infrastrukturkapitel verwiesen, wenn auf die übergreifende Struktur zurückgegriffen wird. Die eingesetzte Infrastruktur ist im Wesentlichen eine Mischung aus Open-Source-Software sowie Eigenentwicklungen am Max-Planck-Institut für Wissenschaftsgeschichte, die als Gemeinschaftswerk einer Reihe von eng zusammenarbeitenden Entwicklern entstanden. Die in dieser Arbeit geschilderten Abläufe der Kommunikation zwischen Fachwissenschaftlerinnen und Fachwissenschaftlern sowie Entwicklerinnen und Entwicklern gelten im Wesentlichen auch hier.

Die Infrastruktur selbst besteht aus einer Reihe von Komponenten, die hier vorgestellt werden. Die Konzeption und die Umsetzung ist ein Gemeinschaftswerk. Die leitenden Ideen für den Aufbau der Umgebung waren, dass es sich um eine Open-Source-Lösung handeln sollte und möglichst viele Standardlösungen für die Software, mehr aber noch für die Datenhaltung eingesetzt werden sollen. Außerdem sollte ein modulares Konzept umgesetzt werden, so dass einzelne Komponenten durch leistungsfähigere oder verallgemeinerte Komponenten ersetzt werden können, wenn dies notwendig und sinnvoll erscheint. Vor allem aber sollte die Lösung so aufgebaut sein, dass flexibel auf die unterschiedlichen Anforderungen in Forschungsprojekten reagiert werden kann. Die vorhergehenden Kapitel haben deutlich gemacht, dass es, bedingt durch die häufig sehr unterschiedlichen Themenfelder, die heterogene Quellenlage und schließlich die hohe Interdisziplinarität der Forschung in den Geisteswissenschaften, für die die Wissenschaftsgeschichte exemplarisch steht, keine monolithische Lösung geben kann, die alle Probleme löst. Insbesondere muss der Aufwand für jeden Einzelfall abgewogen werden. Das bedeutet auch, eine Balance zwischen Lernkurve und Benutzerfreundlichkeit zu finden.[1]

Die Arbeitsteilung innerhalb der gemeinsamen Entwicklungen gestaltete sich wie folgt: Die Oberfläche für die gemeinsame Arbeit mit Python mittels IPython und einer darunter liegenden Infrastruktur für das Filesharing wurde von Malte Vogl realisiert und durch Mittel des Excellence-Clusters Topoi unterstützt (6.1. Die zur Datenauswertung benötigten Komponenten, wie die Anbindung eines Solr-Systems für die Volltextsuche sowie die für diese Arbeit besonders zentrale Arbeitsumgebung zur

[1] Dieses muss jedoch auch Konsequenzen für die Curricula in den verschieden Fächern haben, siehe dazu auch Kapitel 11

Transformation der Daten nach RDF, die Methoden zur Erzeugung der Netzwerkdaten, die Oberflächen für die Abfrage der Daten im Triplestore und die im Laufe der Arbeit immer wieder erwähnten Notizbücher wurden von mir realisiert und sind als Open Source verfügbar (siehe 6.2).

Ein Großteil der Infrastruktur zur Verwaltung der Daten und zu Teilen auch der Publikation dieser Daten basiert auf Django und Django-CMS (6.4). Die für das Kapitel 8 wesentliche Infrastruktur für die Eingabe und Ansicht der Daten des Projektes zur Geschichte der MPG wurden im Wesentlichen von Felix Lange und seinem Team realisiert. Die komplexeren Ansichtsumgebungen für Netzwerke als Erweiterung von Django-CMS stammen vom Autor (6.4.1).

6.1 Die interaktive Umgebung

Da im Forschungsumfeld der Abteilung I des MPIWG eine Reihe von Mitarbeiterinnen und Mitarbeitern mit einem Hintergrund in Physik und Mathematik arbeiteten, war der Gebrauch von digitalen Notizbüchern, wie sie Mathematica und Matlab zur Verfügung stellen, geläufig. Außerdem war Python eine der Sprachen, die in der IT-Gruppe des Institutes häufig eingesetzt wurde. Daher war der Einsatz von Pythonnotizbüchern für die Arbeit mit kurzen Skripten und zur interaktiven Visualisierung eine naheliegende Option. Seitdem durch das Jupyter-Projekt [127] eine prinzipielle Trennung des Kernels, das die Skripte ausführt, und der Notizbuchumgebung realisiert wurde, sind eine Reihe von Erweiterungen entstanden, die es ermöglichen, etwa R, Julia oder auch direkt mit Django interagierende Notizbücher einzusetzen.

6.1.1 Nextcloud

Jupyter wurde in unserer Umgebung an Nextcloud [166] als Filesharing-System angebunden, das die Verwaltung der Dateien übernimmt und Freigaben für andere Benutzer ermöglicht. Die Anbindung erfolgt im wesentlichen durch WebDAV-Mounts [206]. Nextcloud selbst ist eine Open-Source-Umgebung für das Filesharing analog zu kommerziellen Lösungen, wie Dropbox [65], Google Drive [91] oder die iCloud von Apple [116]. Nextcloud ist hierbei eine Abspaltung von ownCloud [179], beide werden zur Zeit parallel zu einander entwickelt. Wir haben uns für Nextcloud entschieden, da sich hier die aus unserer Sicht aktivere Entwicklergruppe gefunden hatte. Prinzipiell lässt sich jedoch unsere Lösung auch mit ownCloud realisieren. Die Nextcloud-Komponente wird zur Verwaltung der Notizbücher und von einzelnen Arbeitsdaten genutzt. Da die Einbindung über das standardisierte WebDAV-Protokoll erfolgt, kann Nextcloud leicht durch andere ähnliche Systeme ersetzt werden. In unserem Kontext wird unter anderem über Alfresco als Alternative diskutiert, insbesondere, falls ein komplexeres Management der Dokumente notwendig sein sollte. Im geschilderten Anwendungsfeld war Nextcloud ausreichend und der Betreuungs- und Installationsaufwand geringer.

6.1.2 Dataverse

Daten, die verarbeitet werden, werden in einem Dataverse-Repositorium [237] abgelegt und können dort publiziert werden. Die Anbindung des Repositoriums an Python und damit an die Notizbücher

erfolgt durch ein Pythonpaket [56]. Dataverse entwickelt vom Institute for Quantitative Social Science (IQSS) der Harvard University ist eine Plattform zur Veröffentlichung von Datensätzen. Datensätze können hierbei mehrere Dateien umfassen. Wir fassen in unseren Anwendungen in der Regel Dateien zu Datensätzen zusammen, die einen unmittelbaren inhaltlichen Bezug haben und in der Regel im gleichen Arbeitsschritt erzeugt werden. Thematisch zusammengehörige Datensätze werden in einem sogenannten Dataverse zusammengefasst. Dataverse erlaubt ein granulares Rechtemanagement in Bezug auf das Anlegen von Datensätzen und Dataverses. Dataverse besitzt ein Versionsmanagement. Zugriffsrechte auf Datensätze lassen sich einschränken, Metadaten zu den Datensätzen sind immer offen. Die grundsätzliche Idee hinter Dataverse ist, dass Datensätze, die im wissenschaftlichen Kontext entstanden sind, in der Regel öffentlich sein sollen und Zugriffsbeschränkung daher die Ausnahme sein sollte. In unserem Falle sind wir aus persönlichkeitsrechtlichen Gründen gezwungen, einen Teil unserer Daten nicht zu veröffentlichen.

6.1.3 Single Sign-on

Systeme, die ein mehrfaches Anmelden eines Benutzers erfordern, haben sich, selbst wenn die Authentifizierung über LDAP [207] erfolgt und daher jeweils nur ein Nutzername und ein Password erforderlich ist, als ein erhebliches Anwendungshemmnis erwiesen. Aufgrund der Sensibilität einzelner Daten sind wir jedoch gezwungen, auch für den lesenden Zugriff auf die Daten eine Anmeldung zu fordern. Dank OpenID [175] müssen Benutzer nur bei der ersten Anmeldung die entsprechenden Freigaben erteilen, danach reicht in der Regel eine einmalige Anmeldung.

6.1.4 Anpassungen für Jupyter/Jupyter-Hub

Eine für unsere Anwendungsszenarien Anforderung an Jupyter war die Möglichkeit, klar definierte Versionen von Notizbüchern anzulegen und stabil vorzuhalten. Von Malte Vogl wurde daher eine Lösung realisiert, die die unterschiedlichen Versionen eines Notizbuches in einer relationalen Datenbank vorhält, so dass diese stabil angesprochen werden können.[2]

6.2 Anwendungen und Anwendungssoftware

Zur Auswertung setzen wir im Folgenden eine Reihe von Standardpaketen sowie einzelne Eigenentwicklungen ein, die weiter unten geschildert werden. Diese sind im Wesentlichen Pythonpakete, außerdem Pakete von R für eine Reihe von statischen Auswertungen.

6.2.1 Netzwerkanalysen

Während der Arbeit mit Netzwerken haben sich zwei Pakete als besonders nützlich erwiesen: *iGraph* und *NetworkX*. Ersteres hat einen C-Kern und ist auch für R erhältlich, während NetworkX vollständig

[2] Zum jetzigen Stand (Dezember 2017) existiert diese Lösung nur als Prototyp. Geplant ist eine Anbindung an das am MPIWG entwickelte System zur Publikation und Erstellung von Open-Access-Publikationen, so dass diese Notizbücher direkt in die Publikationsumgebung eingebunden werden können.

in Python geschrieben ist. Im Wesentlichen benutzen wir diese Pakete für die quantitative Analyse und nur sehr begrenzt für die Visualisierung. Häufig zum Einsatz kommen die in den beiden Paketen implementierten Methoden zur Bestimmung unterschiedlicher Community-Strukturen bzw. Cluster in den Netzwerken.[3]

Hilfreich für eine Übersicht der unterschiedlichen Implementationen und Algorithmen zur Bestimmung von Communitystrukturen und Clustern sind insbesondere auch die zumeist ausführlichen Dokumentationen der Softwarepakete zur Netzwerkanalyse [118] für iGraph oder [160] für NetworkX, die sowohl auf Beispielanwendungen als auch die dahinter stehende Literatur verweisen. Für größere rechenintensive Auswertungen hat sich in der Regel iGraph als deutlich performanter[4] erwiesen. Allerdings haben wir keine expliziten Tests im Hinblick auf die Optimierung von NetworkX durch einen gezielteren Einsatz des Python-C-Compilers *Cython* [49]vorgenommen.

6.2.2 Visualisierungen

Unser ursprünglicher Ansatz war, Visualisierungen in den Notizbüchern selbst durchzuführen. In der Praxis zeigte sich jedoch, dass die Analyse von Graphen mit Standardsoftware wie *Cytoscape* [50] und *Gephi* [87] deutlich effektiver ist. Beide Programme bieten eine Schnittstelle über HTTP an, die es ermöglicht, direkt aus Python und damit auch aus dazugehörigen Notizbüchern Graphen an diese Programme zu übertragen und einzelne Manipulationen durchzuführen. Über die Schnittstelle kann auch das Layout der Graphen geändert werden. Von dieser Möglichkeit machen wir in einer Reihe von Notizbüchern ausführlich Gebrauch. Für komplexere visuelle Analysen übergeben wir daher Graphen direkt an Gephi und Cytoscape. Dieses ermöglicht eine hohe Interaktivität und vermeidet es, bereits existierenden Auswertungsmöglichkeiten noch einmal zu implementieren. Die Kombination von Python und Cytoscape hat sich insbesondere bei der Darstellung der unterschiedlichen Jahresschnitte, die eine zentrale Säule unsere Heuristik darstellen, als äußerst nützlich erwiesen.

Ein Tool, das seine Stärke vor allem beim Vergleich von Graphen und dynamischen Graphen ausspielen kann, ist Visone [252]. Es erlaubt die Animation von zeitabhängigen Graphen und zeigt eine intuitive Visualisierung, die die Veränderungen der Graphen von einer Phase in die andere verdeutlicht. Gleichzeitig kann von Visone direkt die weiter unten beschriebene Modellierungsumgebung RSiena angesprochen werden.

6.2.3 Vom Graphen zum Netzwerk: SPARQLGraph

Zentraler Ansatz unserer Methode ist die Verfügbarmachung von Daten, die semantisch modelliert in einem Triplestore vorliegen, für netzwerkanalytische Methoden. Um den Übergang zu erleichtern, habe ich alle wesentlichen Funktionen zur Erzeugung eines solchen Netzwerkes aus einem Triplestore im Paket *SPARQLGraph* zusammengefasst. SPARQLGraph erlaubt die Erstellung von bipartiten

[3] Siehe auch 5.4.4.

[4] Immer öfter kommt auch *graph-tool* zum Einsatz [96]. Graph-tool unterstützt GPUs und Parallelisierung und ist daher bei Berechnungen von pfadabhängigen Zentralitäten hoch performant. Der Preis dafür ist jedoch die Abhängigkeit von entsprechenden Bibliotheken, die vor allem auf dem Mac unter OS X häufig nur mit einem erheblichen Aufwand installiert werden können.

und mono-modalen Netzwerken direkt mittels SPARQL-Abfragen, es ermöglicht Kanten und Knoten mit entsprechenden Werten aus dem Triplestore zu versehen und schließlich auch neue Daten im Triplestore zu erzeugen [271].

6.2.4 Year-Graph-Format

Für die Analysen von zeitlichen Verläufen ist die Arbeit mit Schnitten unumgänglich. Erweiterungen für iGraph stelle ich in einem Paket *igraphx* zur Verfügung. Methoden dieses Paketes erlauben es, aus Netzwerken mit Jahresangaben, die die Gültigkeit von Knoten und Kanten zu bestimmten Jahren in Form von Zeitintervallen beschreiben, Schnitte für einzelne Jahre zu erstellen. Es ist zusätzlich möglich, Graphen nach bestimmten Kriterien zusammenzufügen. Da die Erstellung von Jahresgraphen in großem Systemen ein unter Umständen rechenaufwändiges Verfahren darstellt, können Jahresgraphen in einem eigenen Format abgespeichert und gelesen werden. Im Wesentlichen ist dies ein gzip-File, das alle Jahresgraphen enthält. *igraphx* enthält Methoden, um diese zu erzeugen und auszuwerten [269].

6.2.5 Ranking von Knoten im zeitlichen Verlauf

Eine der für uns wesentlichen Fragen ist die Veränderung der Stellung von einzelnen Akteuren innerhalb des sozialen Netzwerkes über die Zeit. Hierbei interessieren uns einerseits die Veränderung der absoluten Werte der unterschiedlichen Zentralitäten sowie Veränderungen in der Clusterstruktur, andererseits aber auch die relative Anordnung von Akteuren innerhalb des Clusters. In absoluten Zahlen liegen hierbei die Unterschiede zwischen den einzelnen Akteuren lediglich im Promillebereich, daher bilden wir bei der Auswertung Histogramme mit einer beschränkten Anzahl von Körben und ermitteln den Rang von Personen über diese Körbe.[5]

6.3 Apache Solr

Als Umgebung für die schnelle Suche vor allem in Texten haben wir uns für *Apache Solr* [9] entschieden. Zum Zeitpunkt der Entscheidung war es auf den am MPIWG im Wesentlichen eingesetzten Mac-OS-XSystemen einfacher, Apache-Solr als *Elasticsearch* zu installieren. Daher fiel die Entscheidung auf ersteres.[6]

6.4 Django-CMS

Ergebnisse der Arbeiten sollten auf möglichst einfache Art auch im Intranet verfügbar gemacht werden. Um dieses möglichst einfach zu gestalten liegt der Einsatz eines Content-Management-Systems nahe. Am MPIWG waren zu Beginn der Konzeption der hier in dieser Arbeit geschilderten Umgebungen im wesentlichen *Drupal* [66] und *Zope* [261] im Einsatz. Da ein Großteil der Entwicklungen für

[5] Siehe *igraphx.generateMa_rank* in [269].

[6] Zum Vergleich der beiden Systeme siehe [10].

die Auswertungssoftware selbst pythonbasiert war und langjährige Erfahrungen mit dem ebenfalls pythonbasierten Zope bestanden, fiel die Entscheidung auf *Django* [240] als Framework für die weiteren Entwicklungen, sowie auf die für Django existierende Applikation *Django-CMS* [62] als Content-Management-System. Die Architektur von Django ermöglicht die einfache Anbindung unterschiedlicher SQL-Datenbanken, aber auch von Repositorien wie *Fedora* [80] und insbesondere eine Einbindung von Solr zum Indizieren von mittels Django verwalteten Objekten. Grundsätzlich folgt Django dem Konzept, eines Object-Relation-Mappers, Datenbankstrukturen in Python als Objekte, Klassen und Methoden zur Verfügung zu stellen. Django kann direkt in Pythonnotizbücher eingebunden werden, so dass die in Django verwalteten Objekte dort ausgewertet werden können. Django-CMS wiederum ermöglicht es, über eine Plugin-Architektur Erweiterungen zu entwickeln, die dann über die Editorfunktionen des CMS vom Redakteur der Website in einzelne Seiten eingebunden werden können. Diese sind leichtgewichtige Wrapper-Methoden in Python. Output und Input dieser Methoden werden dann über das Plugin gesteuert.

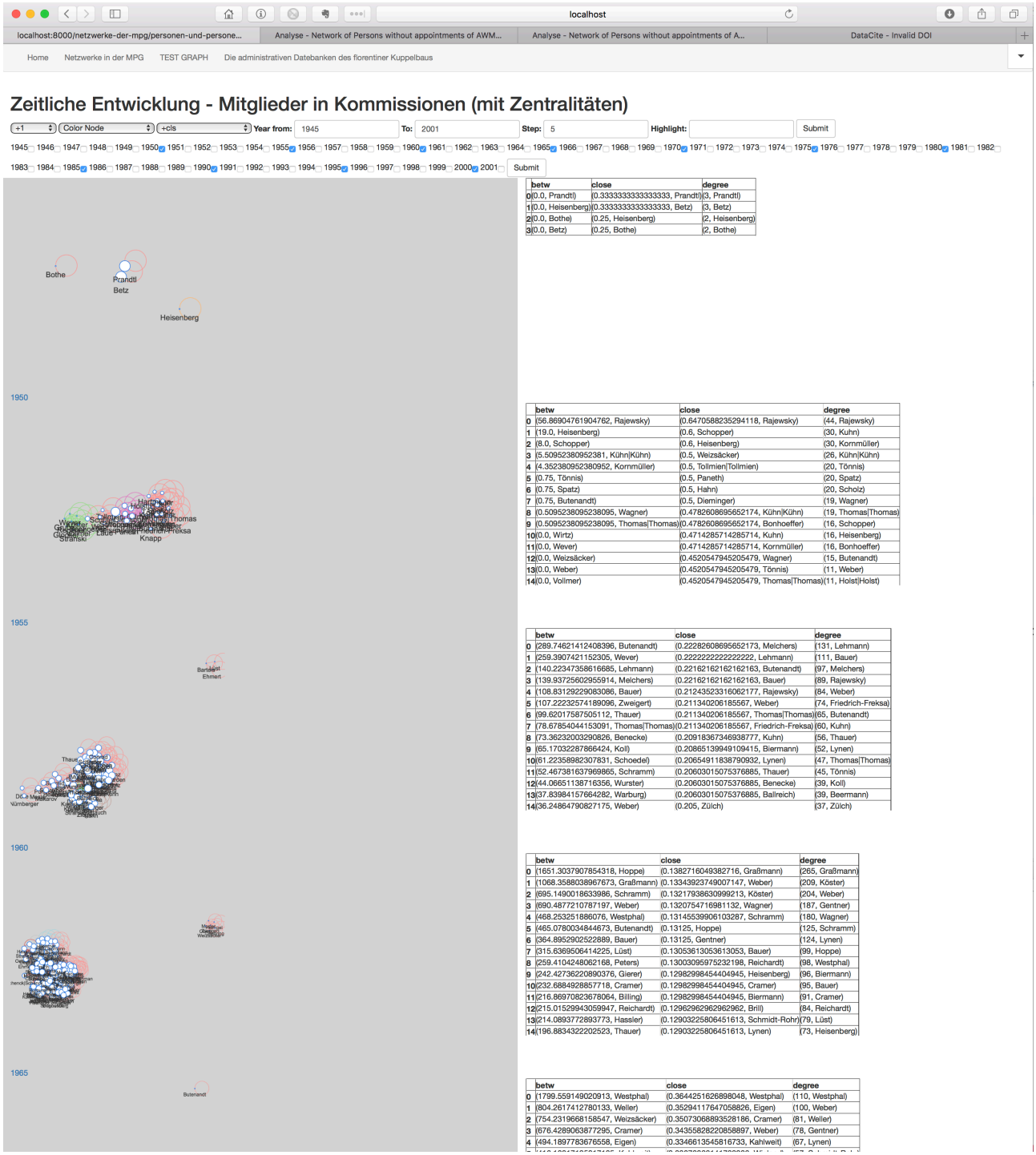
6.4.1 Erweiterungen für Netzwerke und Graphen

Die Webdarstellung von Netzwerken wird in unserer Anwendung durch eine Erweiterung von existierenden Plugins zum Verwalten von Bildern und Dateien realisiert, so dass im CMS gespeicherte Graph- und Year-Graph-Dateien (Abschnitt 6.2.4) direkt in einer Webseite angezeigt werden können. Für die Erstellung von Netzwerken direkt aus dem Triplestore stellt die Erweiterung eine Objektklasse zur Verfügung, die im Wesentlichen die SPARQL-Abfrage als Text speichert und dann daraus die entsprechenden Netzwerke mittels SPARQLGraph (Abschnitt 6.2.3) erstellt und anzeigt. Da dieses Verfahren zeitaufwendig sein kann, werden die Ergebnisse als Graph-ML-Dateien im Dateisystem temporär gespeichert. Auf diese kann dann beim nächsten Aufruf direkt zurückgegriffen werden. Für diese Objekte existiert sowohl die Möglichkeit, die Daten an Gephi oder Cytoscape über die oben geschilderte Schnittstelle zu übergeben, als auch eine Ansicht, die die Dateien direkt in Webseiten einbettet.[7]

Die Abbildung 6.1 zeigt ein Beispiel aus Abschnitt 8.2 für die Darstellung von Jahresgraphen[8] über die Plugins für das Django-CMS. Es ist hier möglich, einzelne Jahre auszuwählen, die Farben von Kanten und Knoten abhängig von den Werten einzelner Attribute zu ändern und einzelne Knoten für alle Jahre hervorzuheben. Außerdem werden Zentralitätsmaße für jedes Jahr berechnet und auf Wunsch mit angezeigt.

[7] Für kleinere Netzwerke, wie die in den Fallbeispielen zur ART und zur Geschichte der MPG, ist die Darstellung über den Browser performant genug, um zumindest einzelne Aspekte der Netzwerke und deren Entwicklung zu untersuchen. Für größere Netzwerke mit mehr als 10.000 Kanten, wie sie sich in den Personennetzwerken des Fallbeispiels zum Archiv des Florentiner Domes ergeben, kommt die Browserdarstellung jedoch an ihre Grenzen und die Streamingfunktion ist wesentlich effektiver.

[8] Details zum Year-Graph-Format und den Tools siehe 6.2.4.



	betw	close	degree
0	(0.0, Prandtl)	(0.3333333333333333, Prandtl)	(3, Prandtl)
1	(0.0, Heisenberg)	(0.3333333333333333, Betz)	(3, Betz)
2	(0.0, Bothe)	(0.25, Heisenberg)	(2, Heisenberg)
3	(0.0, Betz)	(0.25, Bothe)	(2, Bothe)

	betw	close	degree
0	(56.86904761904762, Rajewsky)	(0.6470588235294118, Rajewsky)	(44, Rajewsky)
1	(19.0, Heisenberg)	(0.6, Schopper)	(30, Kuhn)
2	(8.0, Schopper)	(0.6, Heisenberg)	(30, Kornmüller)
3	(5.509523809523809, Kühn|Kühn)	(0.5, Weizsäcker)	(26, Kühn|Kühn)
4	(4.352380952380952, Kornmüller)	(0.5, Tollmien|Tollmien)	(20, Tönnis)
5	(0.75, Tönnis)	(0.5, Paneth)	(20, Spatz)
6	(0.75, Spatz)	(0.5, Hahn)	(20, Scholz)
7	(0.75, Butenandt)	(0.5, Dieminger)	(19, Wagner)
8	(0.5095238095238095, Wagner)	(0.4782608695652174, Kühn|Kühn)	(19, Thomas|Thomas)
9	(0.5095238095238095, Thomas|Thomas)	(0.4782608695652174, Bonhoeffer)	(16, Schopper)
10	(0.0, Wirtz)	(0.4714285714285714, Kuhn)	(16, Heisenberg)
11	(0.0, Wever)	(0.4714285714285714, Kornmüller)	(16, Bonhoeffer)
12	(0.0, Weizsäcker)	(0.4520547945205479, Wagner)	(15, Butenandt)
13	(0.0, Weber)	(0.4520547945205479, Tönnis)	(11, Weber)
14	(0.0, Vollmer)	(0.4520547945205479, Thomas|Thomas)	(11, Holst|Holst)

	betw	close	degree
0	(289.74621412408396, Butenandt)	(0.2228260869565217 3, Melchers)	(131, Lehmann)
1	(259.3907421152305, Wever)	(0.2222222222222222, Lehmann)	(111, Bauer)
2	(140.22347358616685, Lehmann)	(0.22162162162162163, Butenandt)	(97, Melchers)
3	(139.93725602955914, Melchers)	(0.22162162162162163, Bauer)	(89, Rajewsky)
4	(108.83129229083086, Bauer)	(0.21243523316062177, Rajewsky)	(84, Weber)
5	(107.22232574189096, Zweigert)	(0.21134020618556 7, Weber)	(74, Friedrich-Freksa)
6	(99.62017587505112, Thauer)	(0.21134020618556 7, Thomas|Thomas)	(65, Butenandt)
7	(78.67854044153091, Thomas|Thomas)	(0.21134020618556 7, Friedrich-Freksa)	(60, Kuhn)
8	(73.36232003290826, Benecke)	(0.20918367346938777, Kuhn)	(56, Thauer)
9	(65.17032287866424, Koll)	(0.20865139949109415, Biermann)	(52, Lynen)
10	(61.22358982307831, Schoedel)	(0.20654911838790932, Lynen)	(47, Thomas|Thomas)
11	(52.467381637969865, Schramm)	(0.20603015075376885, Thauer)	(45, Tönnis)
12	(44.06651138716356, Wurster)	(0.20603015075376885, Benecke)	(39, Koll)
13	(37.83984157664282, Warburg)	(0.20603015075376885, Ballreich)	(39, Beermann)
14	(36.2486479082717 5, Weber)	(0.205, Zülch)	(37, Zülch)

	betw	close	degree
0	(1651.3037907854318, Hoppe)	(0.1382716049382716, Graßmann)	(265, Graßmann)
1	(1068.3588038967673, Graßmann)	(0.13343923749007147, Weber)	(209, Köster)
2	(695.1490018633986, Schramm)	(0.1321793863099213, Köster)	(204, Weber)
3	(690.4877210787197, Weber)	(0.1320754716981132, Wagner)	(187, Gentner)
4	(468.253251886076, Westphal)	(0.13145539906103287, Schramm)	(180, Wagner)
5	(465.0780034844673, Butenandt)	(0.13125, Hoppe)	(125, Schramm)
6	(364.8952902522889, Bauer)	(0.13125, Gentner)	(124, Lynen)
7	(315.6369506414225, Lüst)	(0.13053613053613053, Bauer)	(99, Hoppe)
8	(259.4104248062168, Peters)	(0.13003095975232198, Reichardt)	(98, Westphal)
9	(242.42736220890376, Gierer)	(0.12982998454404945, Heisenberg)	(96, Biermann)
10	(232.6884928857718, Cramer)	(0.12982998454404945, Cramer)	(95, Bauer)
11	(216.86970823678064, Billing)	(0.12982998454404945, Biermann)	(91, Cramer)
12	(215.01529943059947, Reichardt)	(0.1296296296296962, Brill)	(84, Reichardt)
13	(214.0893772893773, Hassler)	(0.12903225806451613, Schmidt-Rohr)	(79, Lüst)
14	(196.8834322202523, Thauer)	(0.12903225806451613, Lynen)	(73, Heisenberg)

	betw	close	degree
0	(1799.559149020913, Westphal)	(0.3644251626898048, Westphal)	(110, Westphal)
1	(804.2617412780133, Weller)	(0.35294117647058826, Eigen)	(100, Weber)
2	(754.2319668158547, Weizsäcker)	(0.35073068893528186, Cramer)	(81, Weller)
3	(676.4289063877295, Cramer)	(0.34355828220858897, Weber)	(78, Gentner)
4	(494.1897783676558, Eigen)	(0.3346613545816733, Kahlweit)	(67, Lynen)

Abbildung 6.1: Webansicht der Jahresgraphen mit Angabe von Zentralitätsmaßen für die jeweils max. 10 wichtigsten Personen, Beispiel mit Daten von Abschnitt 8.2 ff.

6.5 Statistische Analysen mit R

Insbesondere für die Simulation von wesentlichen Faktoren, die zur Bildung von Netzwerken geführt haben, setzen wir Softwarepakete von R ein. R hat sich, zumindest in der Open-Source-Welt, zum

Standard für statistische Analysen in unterschiedlichsten Anwendungsfeldern entwickelt. Mit *RStudio* und *RStudio-Web* steht für R eine benutzerfreundliche Arbeitsumgebung für R zur Verfügung. Insbesondere auch für die Analyse von Netzwerken liegen eine Reihe von Paketen vor. Wir setzen *iGraph* auch in der R-Variante ein, insbesondere jedoch das zu *statnet* gehörige Paket zur Simulation von Graphen nach den *Exponential Random Graph Models* (*ERGM*).[9] Das Paket erlaubt die Simulation von Graphen auf der Grundlage unterschiedlicher Annahmen.

Tom Snijders stellt mit *RSiena* ein Paket zur Verfügung, um dynamische Entwicklungen von Graphen über mehrere Beobachtungen zu verschiedenen Zeitpunkten zu simulieren und die entsprechenden Faktoren für ihre Genese abzuschätzen.[10] Es stehen sowohl Modelle abhängig von Werten der Knoten als auch für Wechselbeziehungen zur Verfügung. RSiena lässt sich direkt aus dem Visualisierungsprogramm Visone für Graphen aufrufen.

Da wir es später häufiger mit in der Statistik üblichen Darstellungen zu tun haben werden, die sich als Ergebnis unterschiedlicher Auswertungen mit R ergeben (wie in Tabelle 6.1) geben wir hier eine kurze Erklärung, wie diese zu interpretieren sind. Jede Zeile steht für einen Effekt. Die dahinter stehende Zahl gibt die Stärke des Effektes, der Wert in Klammern (manchmal auch in einer weiteren Spalte) die Standardabweichung und schließlich die Anzahl der „*“ in der Regel die Güte des Wertes an. Die Stärke des Effektes ist hierbei der Logarithmus der Wahrscheinlichkeit dafür, dass sich Paare in unserem Netzwerk bilden, wenn der entsprechende Effekt zutrifft. Negative Werte stehen damit für abschwächende Effekte und positive Werte für verstärkende Effekte. Die mit den Sternen bezeichneten p-Werte entsprechen den in der Statistik üblichen Signifikanzniveaus, also der Wahrscheinlichkeit eine Nullhypothese fälschlich abzulehnen. In der Statistik werden üblicherweise Werte < 0.05 als signifikant, < 0.01 als sehr signifikant und < 0.001 als extrem signifikant bezeichnet. Bei Simulationen in der Sozialen Netzwerkanalyse gelten üblicherweise schon Werte kleiner als < 0.1 als signifikant.

[9] Siehe 5.6.

[10] Zu SIENA-Modellen siehe auch Abschnitt 5.7.

	Dependent variable:
	grN
edges	−5.664***
	(0.173)
nodematch.disc	0.173***
	(0.025)
nodematch.nat	0.645***
	(0.041)
gwesp.fixed.0	1.732***
	(0.170)
Akaike Inf. Crit.	18,808.780
Bayesian Inf. Crit.	18,851.270
Note:	*p<0.1; **p<0.05; ***p<0.01

Tabelle 6.1: Beispiel der Ergebnisse nach einer Simulation mit ERGM

6.6 Triplestores

In Kapitel 4 hatten wir grundsätzlich die Bedeutung von Graphen für unsere Arbeiten vorgestellt und besonders in Abschnitt 4.1.3 auf die Rolle von Triplestores hingewiesen. Hier soll nun die konkrete Realisierung vorgestellt werden.[11] Für die Verwaltung von RDF-Tripeln mit ihren Kontexten – also de facto Quadrupeln – stehen mittlerweile verschiedenste Lösungen als Open-Source zur Verfügung. Wir haben uns aus pragmatischen Gründen für Blazegraph [55, 29], das vormalige *bigdata*, entschieden. Als Java basierte Anwendung ist es leicht zu installieren und zu betreiben, der Datenimport ist hoch performant und es bietet eine Reihe von einfach anzusprechenden hilfreichen Erweiterungen von SPARQL, insbesondere eine indexbasierte Suche auf Volltexten sowie Erweiterungen zur schnel-

[11] Für ein großes Projekt zur Erfassung von Metadaten und Digitalisaten von Manuskripten aus dem islamischen Kulturkreis, die Islamic Scientific Manuscript Initiative (ISMI) [238], hielten wir es bereits in den frühen 2000er Jahren für notwendig, ein flexibles Datenbanksystem zu entwickeln, das freie Assoziationen von Attributen und Relationen ermöglicht. In Form von OpenMind wurde es in vielen unterschiedlichen Phasen entwickelt, an denen eine Reihe von Entwicklern beteiligt waren, zu nennen sind hier Kai Stallmann, Jorge Urzua und Robert Casties. Letztendlich nahm unsere Eigenentwicklung eine Reihe von Funktionen vorweg, die nun vor allem durch Neo4J abgedeckt werden. Auch erreichte unser System niemals die Performance, die Neo4J bietet, wobei das vor allem von Jorge Urzua für OpenMind entwickelte Eingabesystem jedoch auch heute noch im Einsatz ist und die strukturierte und konsistente Eingabe von Daten vor allem durch die Wissenschaftler selbst erleichtert. Auf der Abfrageseite reicht unser System sicher in keiner Weise an die Performance von Neo4J heran.

len Erzeugung von Teilgraphen durch Angabe von Ressourcen als Grenzbedingungen, für die dann verbindende Pfade gefunden werden. SPARQL Property-Paths [223] werden performant unterstützt. Vor allem aber steht mit *metaphactory* von *metaphacts* [205] und der darauf aufbauenden für Forschungszwecke angepassten Oberfläche *ResearchSpace* [205] eine Umgebung zur Verfügung, die den Umgang mit Triplestores auch mit geringen SPARQL-Kenntnissen ermöglicht.[12]. In unseren Fallbeispielen wird *ResearchSpace* nicht direkt eingesetzt, sondern es werden lediglich Funktionalitäten von *metaphactory* genutzt.

Metaphactory bietet ein einfaches Templating-System an. In HTML-Seiten werden hier auf *ReactJS* [197] basierende Elemente eingebaut, die SPARQL-Abfragen auch in dynamischer Form enthalten und dann die Ergebnisse als Tabellen, Graphen oder Netzwerke darstellen. Auch Suchabfragen lassen sich in *metaphactory* direkt realisieren. Komplexere Abfragen, solange sie im wesentlichen *CRM*-basiert sind, werden dann durch ResearchSpace realisiert. ResearchSpace ermöglicht auch die Eingabe von Daten. Dieses wird im wesentlichen durch den Einsatz von *LDP* (siehe Abschnitt 4.2.3) realisiert.

Neue Templates können ähnlich wie bei Wikis direkt über die Weboberfläche angelegt werden, so dass datenbasierte Präsentationen in Form von Webseiten schnell ohne größere Progammierkenntnisse realisiert werden können. SPARQL Kenntnisse sind jedoch erforderlich.

6.7 Jenkins-Workflow

Für einzelne Abläufe benötigen wir automatisierte Skripte, die regelmäßig neue Versionen unserer Datenbanken erstellen. Insbesondere gilt dieses für die aus den SQL-basierten Daten erstellten abgeleiteten Repräsentationen von Daten im Triplestore in Kapitel 8. Als ein einfaches und übersichtlich gestaltetes Hilfsmittel hat sich hier das ursprüngliche für *continous integration* entwickelte *Apache-Jenkins* [123] erwiesen. Skripte können hier einfach auch auf verteilten Systemen ausgeführt und die Ergebnisse verlässlich überwacht werden. Die Abbildung 6.2 zeigt beispielhaft einen Ausschnitt der Übersicht über die Skripte, die zum täglichen Abgleich des Triplestores mit den Daten des GMPG-Projektes genutzt werden.

6.8 Authorities, Normdaten, Linked Open Data

In den einzelnen Fallstudien greifen wir auf Normdaten aus verschiedenen Kontexten zurück. Überwiegend sind dieses in unseren Anwendungsbeispielen Daten aus der *Gemeinsamen Normdatei* (*GND*) [61], die die Deutsche Nationalbibliothek zur Verfügung stellt, sowie Daten aus *Wikidata* [264]. Diese liegen als RDF-Datensätze vor[13] und sind direkt im Netz über SPARQL-Endpunkte abfragbar.[14] Einzelne Datensätze werden in der Regel mit stabilen URL als *Linked Open Data* zur Verfügung gestellt. In den einzelnen Anwendungsbeispielen werden wir jeweils genauer auf die Funktion der

[12] https://github.com/researchspace/researchspace

[13] Für die GND siehe [60], für Wikidata siehe [119].

[14] Für die GND siehe [224], für Wikidata siehe [266].

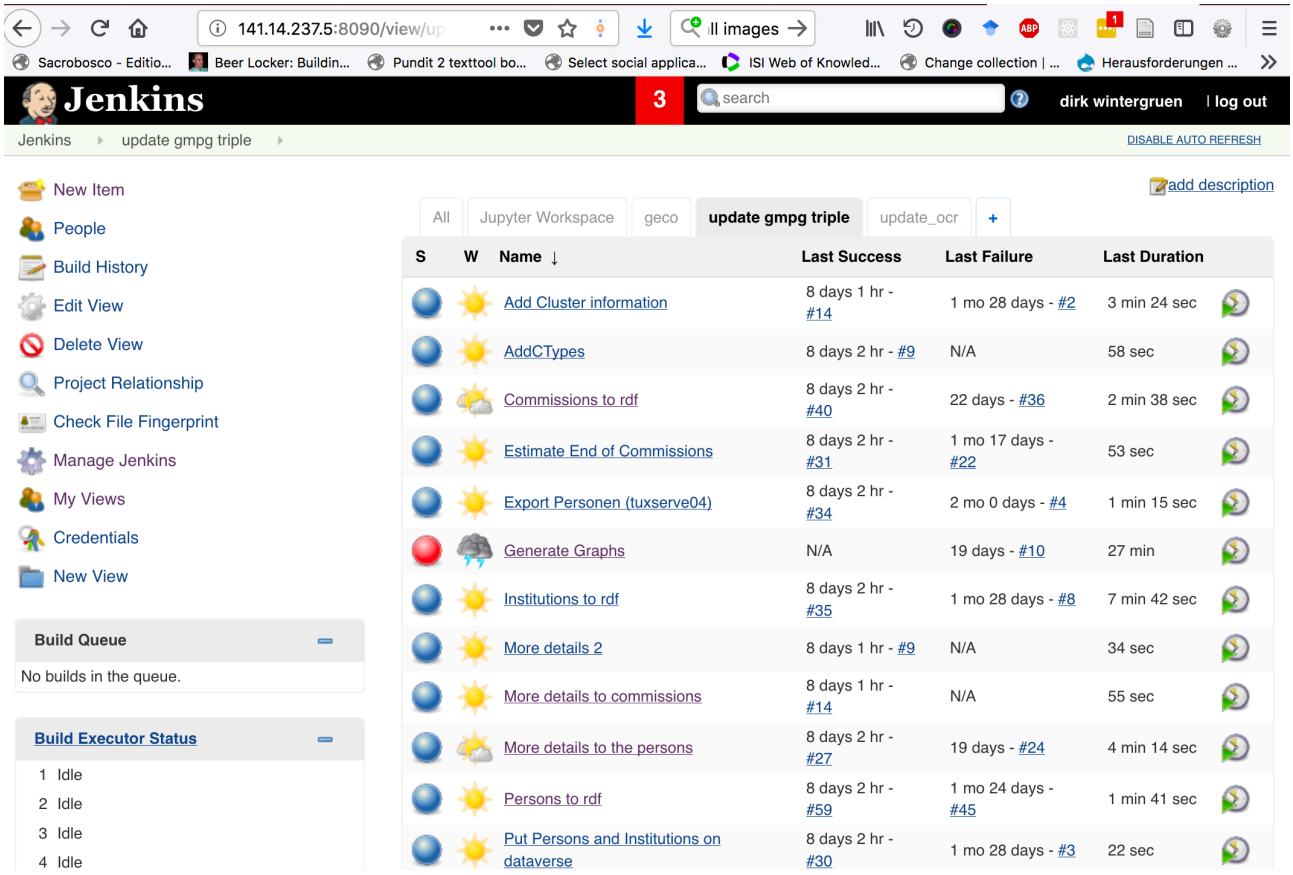


Abbildung 6.2: Jenkins: Übersicht über die Skripte zum regelmäßigen Update des Triplestore

einzelnen Datenbestände und deren Einsatzmöglichkeiten eingehen.[15] Hier sollen nur kurz das allgemeine Vorgehen im Umgang mit diesen Daten und die gemachten Erfahrungen zusammengefasst werden. Als Grundregel gilt leider immer noch, insbesondere für die Daten, die als RDF in irgendeiner Form bereitgestellt werden, dass es eines manchmal doch erheblichen Aufwandes bedarf, die entsprechenden – mittlerweile doch reichhaltig – vorhandenen Datensätze zu finden. Die Benutzung von bereitgestellten Endpunkten im Netz ist immer noch oft mühsam, da die Performanz zu wünschen übrig lässt. Selbst wenn Einzelsuchen möglich sind, gelingen Massenabfragen häufig nicht oder nur mit doch erheblichen Wartezeiten. Die effektivste Lösung ist in vielen Fällen der Download der RDF-Dateien und die Bereitstellung in einem eigenen lokalen Triplestore.[16] Mit Referenzen im Kontext von *linked open data* waren die Erfahrungen ebenfalls gemischt. Wikidata mit seinen Referenzen zu Wikipedia hat sich als zuverlässig erwiesen, dieses gilt auch für die GND und VIAF, während bei Referenzen auf WordNet immer wieder Probleme auftraten.

Trotzdem hat sich gerade WordNet als hilfreiche Ressource erwiesen. In Abschnitt 9.3.4 zeigen wir, wie mit Hilfe von standardisierten Begriffen über eine Verbindung zu WordNet [229, 153, 177]

[15] Abschnitte 7.3.2 und 8.3.3.

[16] Auch hier zeigt sich, dass eine engere Abstimmung sinnvoll wäre. In persönlichen Gesprächen mit vielen Kolleginnen und Kollegen kam häufig die Reaktion: Haben wir auch so gemacht! Eine föderierte Architektur, die diese Versuche zusammenfasst, wäre vielleicht ein hilfreicher Ansatz.

als *linked open data*-Ressource für englischsprachige Konzepte und Wortstämme Ansätze für multilinguale Zugänge zu Datenbanken realisiert werden können. Im Detail war dieses nicht immer unproblematisch zu realisieren, da sich die URLs, die es ermöglichen, die Konzepte maschinenlesbar auszuwerten, über die Zeit verändert hatten. Jedoch möchte ich betonen, dass die Bereitstellung der Daten des WordNet in allen seinen Versionen zum Download in verschiedenen Formen eine große Hilfe für viele Arbeiten mit textbasierten Problemen ist.[17] Es ist sehr bedauerlich, dass für andere Sprachen als das Englische der Zugang zu entsprechenden Datensätzen nicht so einfach möglich ist. Experimente mit multilingualen Zugängen zu Datenbanken würden sich dadurch erheblich vereinfachen.

[17] Und dafür vielen Dank an die Verfasser und Betreuer dieser Seiten.

Kapitel 7

Die Netzwerke der Allgemeinen Relativitätstheorie

Das folgende Kapitel entstand aus einer Gemeinschaftsarbeit von Roberto Lalli und dem Autor. Die Grundlage für die Arbeit sind die von Roberto Lalli zusammengetragenen Daten und seine historische Interpretation der Ergebnisse. Von meiner Seite stammt die konkrete Implementation der Verfahren und Methoden, die hier im Folgenden näher beschrieben werden. In gemeinsamer Arbeit sind die hinter dem Netzwerk stehenden Annahmen über Kooperationsbeziehungen und insbesondere deren Gewichtung entstanden. Die zentrale Frage, der wir in diesem Kapitel nachgehen werden, ist, ob und wie sich die strukturellen Veränderungen des Feldes der Forschung an der *Allgemeinen Relativitätstheorie (ART)* in den Akteursnetzwerken über den Zeitverlauf widerspiegeln. Die Ausgangsthese ist, dass sich nach einer längeren Verzögerungsphase in den 1960er Jahren die Allgemeine Relativitätstheorie als eigenständige Disziplin etablierte. Diese Phase wird in der Literatur als die Zeit der Renaissance der ART bezeichnet [30, 136], während die Phase davor nach Jean Eisenstaedt [71] als Low-Water-Mark-Periode der ART charakterisiert wird. Kriterien für die Disziplinenbildung sind hierbei u.a. die Etablierung von Institutionen und stabilen Forschergruppen, die sich mit diesem Thema im Zentrum ihrer Forschung beschäftigen, sowie der Anstieg der Zahl der Publikationen in diesem Feld – insbesondere in eigenen spezifischen Zeitschriftenreihen. Wir wollen an dieser Stelle diesen wissenschaftshistorischen Erklärungsansatz nicht im Detail diskutieren oder kritisieren. Im Zentrum dieses Abschnittes steht eine Darstellung unseres methodischen Vorgehens, um uns dieser Fragestellung mit Hilfe digitaler Auswertungsmethoden anzunähern. Das Kapitel beleuchtet hierbei insbesondere den Prozess der gemeinsamen Arbeit, das Wechselspiel von Analyse, Visualisierung, Entwicklung und Diskussion, das zur formalen Beschreibung des Netzwerkes führte und damit zur Möglichkeit, die uns vorliegenden Daten mit mathematischen Methoden zu untersuchen. Es beleuchtet zugleich die Probleme der Datenlage im Hinblick auf die Validität der Ergebnisse.

7.1 Voraussetzungen – methodische Probleme

Wir haben diese Fallstudie gewählt, da die unterliegenden Entwicklungen der Geschichte der Allgemeinen Relativitätstheorie bereits aus wissenschaftshistorischer Sicht gut erforscht wurden und wir somit Ergebnisse, die wir mittels quantitativer Methoden erhalten, systematisch mit dem Forschungsstand vergleichen können. Insbesondere ist ein wesentlicher Teil der Datengrundlage, das Beziehungsgeflecht der Hauptakteure, weitestgehend verlässlich überprüft. Problematischer sind die Anteile der Studie, die auf der Analyse von Zitationsbeziehungen beruhen. Hier liegen uns nur Daten kommerzieller Anbieter[1] vor. Einige Fehler sind hier offensichtlich, wie Fehler in Schreibweisen von Namen. Insbesondere aber sind die Personennamen im Web of Science nicht vereinheitlich und nicht mit Normdaten verbunden, zumindest wenn es sich um weit zurückreichende Einträge handelt. So haben wir noch keinen abschließenden Eindruck über die Vollständigkeit der Daten in den Frühphasen der ART gewinnen können. Mit Sicherheit fehlen Beiträge aus dem nicht-anglophonen Bereich – insbesondere Veröffentlichungen in russischer Sprache.[2] Daher besteht hier noch stärker als bei der Untersuchung der sozialen Struktur die Gefahr systematischer Fehler. Inwieweit die Schwerpunktsetzung bei der Datensammlung, geprägt durch die Relevanz von Publikationen aus heutiger Sicht, die Ausprägung von Themenfeldern in der historischen Analyse verfälscht, können wir zum jetzigen Zeitpunkt noch nicht abschließend beurteilen. Dies gilt insbesondere für die von uns beobachtete gute Übereinstimmung der zeitlichen Entwicklung, die wir aus der wissenschaftshistorischen Aufarbeitung des Feldes kennen, mit den Ergebnissen der Kozitationsanalyse. Wir sind uns hier des Risikos bewusst, dass die tatsächliche historische Situation von Auswahlkriterien der Datenbank wenn nicht überlagert, so doch zumindest beeinflusst wird.

Auch bei der Konstruktion und Bewertung sind methodische Probleme zu berücksichtigen, die die Interpretation der Ergebnisse beeinflussen. In unserem Ansatz sind wir von den Arbeiten von Tom Snijders und Emmanuel Lazega[3] – sowie insbesondere von Elisa Bellotti[4] über Multilevel-Netzwerke zur Untersuchung von Strukturen in der Forschung inspiriert worden. Unsere Fallstudie stellt uns jedoch vor Herausforderungen, die bei einer direkten Übertragung der dort vorgestellten Methoden zu Problemen führen. Wir untersuchen eine Zeitperiode, die einerseits zwar überschaubar andererseits jedoch so lang ist, dass sie große historische Umbrüche umfasst. Quantitative Untersuchungen in den Übergangsphasen mit einfachen Modellen im Kontext von SIENA[5] und ERGM[6] sind daher problematisch. Die Aufteilung der Modellrechnungen in eine Phase vor und eine nach dem 2. Weltkrieg liegt nahe, ist aber offensichtlich unbefriedigend, da gerade die Rückwirkungen der Phase vor dem 2. Weltkrieg und die Umbruchphase danach bis in den Kalten Krieg eine der wesentlichen Fragen an

[1] Im Wesentlichen sind das Daten des Web of Science, die von Clarivate Analytics angeboten werden.

[2] Dies ist insbesondere deshalb kritisch, da nur wenige zusätzliche Publikationen die Rolle der Personen, die Vermittler zwischen Ost und West sind, deutlich verändern können, denn hier sind zurzeit nur wenige gemeinsame Publikationen bekannt. Diese Personen als Mittler haben daher eine hohe Betweenness (siehe 5.3.2), die sich augenblicklich stark verändert, wenn neue Publikationen identifiziert werden können.

[3] Ein gute Zusammenfassung ihrer Ansätze findet sich in [138].

[4] Ursprünglich in [20] und dann ausführlicher in [19].

[5] Siehe Abschnitt 5.7.

[6] Siehe Abschnitt 5.6.

die Untersuchung der Netzwerkdynamik sind.

Dies führt uns zu grundsätzlichen methodischen Problemen bei der Bewertung der Netzwerkstrukturen. So ist insbesondere die Frage der Gewichtung von Beziehungen über den Untersuchungszeitraum und ihrer Interpretation problematisch. Die Bedeutung von nationalen Bindungen hat sich in den Zeiträumen vor dem 2. Weltkrieg über die Kriegsphase und schließlich bis in den Kalten Krieg hinein grundlegend verändert. Disziplinen und disziplinäre Zuordnungen haben sich verschoben und die Rolle der Universitäten und Forschungseinrichtungen unterliegt großen Veränderungen. Diese Umbrüche bestimmen einen erheblichen Teil des Untersuchungszeitraumes. Schließlich wächst die Anzahl der beteiligten Personen in der Hochphase des Feldes dramatisch an. Außerdem können wir keine quantitativen Größen definieren, wie etwa im Falle von Bellottis Studien [19] Geldströme, die es ermöglichen, Netzwerkstrukturen direkt mit messbaren Größen zu korrelieren. Unsere entscheidenden Größen sind topologischer Natur, wie die Frage nach der Stellung von Einzelpersonen im Netz und die Bildung von Gruppen und Institutionen. Die im Allgemeinen als für die Etablierung von stabilen Strukturen als wesentlicher Faktor angesehene Entwicklung einer großen, zusammenhängenden, relativ stabilen *Komponente*[7] des Netzwerkes ist hierbei auch für uns von großem Gewicht. Dafür ist jedoch die Klärung der Frage nach ihrer inneren Struktur und insbesondere die Frage der Geschichte der größten Komponenten aus jeweils rückblickender Sicht nicht unproblematisch. Einfach gefragt: Ist die größte Komponente von 1980 dieselbe wie die von 1960 oder 1940 oder ist diese aus einer ganz anderen Kerngruppe entstanden? Da sich die Knoten in den Komponenten durch Ausscheiden und Hinzukommen von Personen in das Netzwerk über die Zeit ändern, ist die Frage, wie die Ähnlichkeit von Komponenten über eine längere historische Entwicklungsphase definiert werden soll, nicht trivial. Auch die Länge der Nachwirkung von aufgebauten Bindungen (oder eventuell ihr Vorwirken) ist nicht unmittelbar zu beurteilen: Wann endet eine Bindung in einem Netzwerk, wie lange hallt ihr Einfluss nach?

Dies machte aus unserer Sicht jedoch umgekehrt diese Fallstudie gerade zu einem für die historische Arbeit mit Netzwerken im Hinblick auf die Evaluation und Erprobung der Methodik vielversprechenden Anwendungsszenario. Das vorliegende Kapitel bietet nicht den Raum, alle Schritte bei der Arbeit an diesen Netzwerken detailliert zu beschreiben, und so kann nur das Wesentliche herausgegriffen werden. Jedoch soll die Darstellung im Folgenden zeigen, welche Bedeutung der als zentral hinter der gesamten Arbeit stehende Ansatz einer Verknüpfung von quantitativer Datenanalyse und semantischer Modellierung, die Möglichkeit interaktiv zu arbeiten und schließlich die Netzwerkanalyse für solche Studien haben. Sie zeigt den prozessoralen Charakter der Arbeit mit solchen Tools, der die Offenheit bewahrt, auf neue Anforderungen kontinuierlich zu reagieren.

Die Arbeiten begannen mit mehr oder weniger naiven Visualisierungen der Netzwerkstrukturen und wurden in jeder Iterationsrunde verändert. Sie sind immer noch einem kontinuierlichen Anpassungsprozess unterworfen – neue Daten und neue Ansichten erforderten und erfordern immer wieder eine radikale Umbewertung. Wir haben längst nicht auf alle methodischen Fragen eine Antwort gefunden. So sind die geschilderten Ergebnisse, wie jede historische Arbeit, vor dem Hintergrund unserer heutigen Wissensgrundlage zu interpretieren. Dafür ist die folgenden Fallstudie ein aus unserer Sicht

[7] Siehe 5.3.2 zum Begriff der Komponente.

überzeugendes Beispiel. Sie zeigt die Rolle von quantitativen Methoden als heuristisches Instrument für die Geisteswissenschaft, dem Leitmotiv dieser Arbeit. Wie bei allen folgenden Fallstudien geht es uns hier um eine nachvollziehbare Darstellung der Methodik sowohl von Seiten der Informatik als auch und insbesondere von Seiten der Geisteswissenschaften. Sie steht damit einer erwünschten kritischen Diskussion offen.

7.2 Konstruktion und Struktur der sozialen Netzwerke

Das Netzwerk, das wir im Folgenden betrachten werden, ist ein Multilevel-Netzwerk der folgenden Struktur: Es besteht aus zwei Hauptbestandteilen, bei denen wir auf der einen Seite Beziehungen haben, die aus einem sozialen Kontakt entstehen, also ein soziales Interaktionsnetzwerk. Auf der anderen Seite haben wir Beziehungen, von denen eine methodische Übereinstimmung oder ein Einfluss auf die Wahl von Methoden oder auf Forschungsfelder ausgehen können, die aber nicht unmittelbar soziale Kontakte implizieren bzw. die wir nicht auf Grund der uns vorliegenden Quellen verifizieren können. So enthält die erste Gruppe von Teilnetzen Personen und Institutionen als Knoten. Innerhalb dieser Teilnetze beschreiben wir ein *ungerichtetes monomodales* Netzwerk, das die gleichberechtigte Kooperation zwischen Wissenschaftlern beschreibt, ein gerichtetes Netzwerk von Schüler-Lehrer- oder Abhängigkeitsverhältnissen, sowie ein *bipartites* Teilnetz, das Forscher und Institutionen verbindet. Kooperationen zwischen Institutionen sind im Prinzip ebenfalls möglich. Diese sind aber durch unsere Datenlage noch nicht ausreichend gestützt.

Der zweite Hauptbestandteil des Netzwerkes sind Beziehungen, die wir aufgrund inhaltlicher Übereinstimmung bzw. als gemeinsame Grundlage für einen möglichen Methodenwandel annehmen. Dies ist die Zugehörigkeit zu einem Forschungsfeld bzw. einer Subdisziplin, die zu einem Methodenaustausch geführt haben könnten, sowie Beziehungen zwischen Personen, die sich aus der Analyse von Kozitationen ergeben. In diese Kategorie von Beziehungen fällt hier auch die Nationalität.[8] Wir haben also auch auf dieser Ebene mehrere bipartite Netzwerke: Zum jetzigen Zeitpunkt sind dies Personen und Nationen sowie Personen und Forschungsfelder. Unsere Forschungsfragen beziehen sich im Wesentlichen auf das Verhältnis zwischen diesen beiden Hauptbestandteilen und deren gegenseitige Abhängigkeit.

Systematisch untersuchen wir zunächst die Struktur des ersten Bestandteils, also den im engeren Sinne sozialen Zusammenhang, und berücksichtigen zunächst den zweiten Aspekt nur unter dem Gesichtspunkt von möglichen Korrelationen. Wir sehen hier Zuordnungen zu Nationen, Disziplinen und Forschungsfeldern als abhängige Variablen des sozialen Netzwerkes. Das vollständige Netzwerk stellt sich in der von Elisa Bellotti und anderen in [19, p.219] [9] angegeben Klassifikation als ein *MMMC*-Problem (für den ersten Teil unseres Netzes) und ein *MNA*-Problem (für den zweiten Teil des Netzes) dar. In 7.5 gehen wir kurz auf erweiterte Ansätze zum Umgang mit Forschungsprogrammen und Ansätze zur Multilevel-Analyse dieser Strukturen ein.

[8] Da wir nicht davon ausgehen, dass Nationalität wie im im Falle von Institutionen unseres Netzwerkes zu einem direkten Austausch geführt haben könnten.

[9] Siehe auch Kapitel 5.8.

Wir folgen hier in Teilen den von Ales Ziberna und Lazega [277] gemachten Vorschlägen für eine vergleichende Untersuchung vor allem der Block- und Clusterstrukturen der Netzwerke unter verschiedenen Voraussetzungen. Da wir zunächst keine Beziehungen zwischen den Institutionen annehmen, vergleichen wir jedoch nur die Ergebnisse aus der Analyse des *erweiterten Netzwerkes*,[10] das durch das Einfügen von zusätzlichen Kanten entsteht, die sich aus dem *bipartiten* Institutionennetzwerk ergeben, mit den Ergebnissen eines vollständigen *Multilevel-Ansatzes*. Letztendlich gehen wir davon aus, dass nur ein Multilevel-Ansatz die vollständige Dynamik des Prozesses hinreichend erklären kann, da zwei der wesentlichen Faktoren, die zur Ausprägung des Feldes geführt haben – eine einsetzende Institutionalisierung und die unmittelbare Kooperation der beteiligten Forscher –, in enger Wechselbeziehung stehende Netzwerke darstellen. Die Beschreibung der Institutionalisierung alleine erfordert, wie bereits oben beschrieben, mindestens ein bipartites Netzwerk zwischen Personen und Institutionen sowie ein Netzwerk von Institutionen untereinander.

Von besonderem Gewicht sind bei unseren Untersuchungen die zeitlichen Entwicklungen des Netzwerkes. Dies wirft zusätzlich zur statistischen Gewichtung der Relationen im Netz das Problem einer Entwicklungsdynamik und das damit verbundene Problem der Skalen auf, die adäquat sind, um unterschiedliche Fälle vergleichbar machen zu können.

In den Arbeiten von Bettencourt und Kaiser [22] wird dies näher betrachtet. Ihr wesentliches Ergebnis ist, dass die Anzahl der beteiligten Personen ein besseres Kriterium für Entwicklungsdynamik ist als die Zeit als laufender Parameter.[11] Dies wiederum setzt voraus, dass die Korrelation zwischen Personen und Zeit zumindest monoton ist. Wir können das dadurch sicherstellen, dass wir annehmen, dass einmal aufgebaute Bindungen im Gedächtnis des Gesamtsystems erhalten bleiben. Wir betrachten diese Annahme zumindest als kritisch und untersuchen daher Netzwerke mit verschiedenen zeitabhängigen Gewichtungsfaktoren, wobei wir einen Zusammenhang mit unterschiedlichen Konstanten c als *Nachhallparameter* sowie unterschiedlichen Gewichten e_{ij} für unterschiedliche Beziehungen annehmen.[12]

$$f_{ij}(t, t_0(i)) = e_{ij}\delta_{t,t_0(i)} \text{ mit } \delta_{t,t_0(i)} \begin{cases} 1 & \text{wenn } t + t_0 < c \\ 0 & \text{sonst.} \end{cases} \tag{7.1}$$

Für die Gewichte der Beziehungen zwischen Personen setzen wir zusätzlich eine Kurve in der Form einer Eulerschen Fehlerfunktion über die Zeit an, so dass Bindungen stärker gewichtet werden, wenn sie länger andauern. Außerdem nehmen wir an, dass der Einfluss von Personen über die Form eines institutionellen Gedächtnisses nachwirkt, d.h. also, auch wenn sich Anwesenheiten nicht überlagern, setzen wir einen Einfluss an, der gegen 0 geht, wenn die Zeiträume entsprechend größer werden.

Es gilt also:

$$I(\Delta_t) = \begin{cases} \frac{m_- + f_{erf}\left(\frac{\Delta_t}{d_-}\right)}{k_-} & \text{wenn } \delta_0 < 0 \\ \frac{m_+ + f_{erf}\left(\frac{\Delta_t}{d_+}\right)}{k_+} & \delta_0 \geq 0 \end{cases} \tag{7.2}$$

[10] Siehe auch Kapitel 5.8.

[11] Siehe 5.9.

[12] Siehe auch Kapitel 8.5.1.

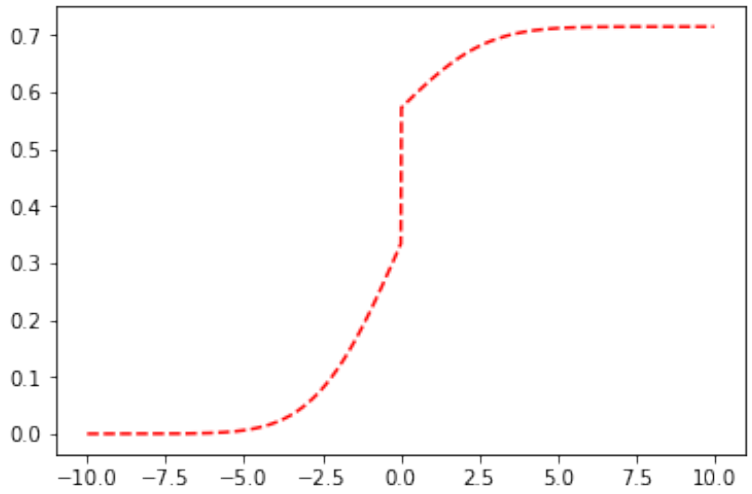


Abbildung 7.1: Dämpfungsfunktion für $d_+ = d_- = 3, k_- = 3, k_+ = 7, m_- = 1, m_+ = 4$

$$f_{erf}(x) = \int_x^\infty \frac{2}{\sqrt{\pi}} e^{-r^2}\, dt \tag{7.3}$$

Die Konstanten $m_-, m_+, d_-, d_+, k_-, k_+$ bestimmen hierbei das Dämpfungsverhalten. Wir nehmen jeweils an: $1 \leq m_- \leq m_+,\ 1 \leq k_- \leq k_+$, sowie $d_+ = d_- \geq 1$.

Unsere Fragen an das Netz sind zunächst klassische Fragen an die Rolle von Einzelpersonen in sozialen Netzen. Welches sind die Personen, die in unserem System zu welchem Zeitpunkt den größten Einfluss ausüben, und welche verallgemeinerbaren Strukturen können wir erkennen? Die Frage nach dem Maß von Einfluss ist hier vor allem auch eine historische Frage. Die zweite Frage ist inhaltlich-systematischer Art die Frage nach dem Zeitpunkt des Aufkommens neuer Forschungsfelder und der Rolle dieser Forschungsfelder. Der ersten Frage gehen wir im Folgenden ausführlicher nach. Die zweite Frage können wir nur anreißen, weil bereits bei den ersten Versuchen, die Rolle der Forschungsfelder im Rahmen der Netzwerktheorie zu behandeln, offensichtlich wurde, dass eine genauere Untersuchung der historischen Prozesse notwendig ist, um eine Zuordnung von Personen und Institutionen zu Forschungsfeldern, Forschungsagenden und Disziplinen sowohl im historischen Kontext als auch aus der Rückschau zu erreichen, die präzise genug sind, so dass sie mit Hilfe von Modellen wie ERGM oder SIENA behandelt werden können.

7.3 Die Datenbasis und ihre Modellierung

Die Ausgangslage für die Arbeit mit den Daten war eine Filemaker-Datenbank, die auch derzeit noch für die Dateneingabe und -korrektur benutzt wird. Als Konsequenz muss auch für die Zukunft ein Import bzw. eine Transformation der Daten in die für die Analysen genutzte Datenstruktur in RDF sicher gestellt sein. Da die Eingaben der Grunddaten nur in der Filemaker-Datenbank erfolgen, ist eine Synchronisation der Daten nicht erforderlich.[13]

[13] Eine Infrastruktur, die auch die direkte Eingabe in den Triplestore ermöglicht, soll jedoch mittelfristig ebenfalls zur Verfügung gestellt werden. Siehe auch 6.6.

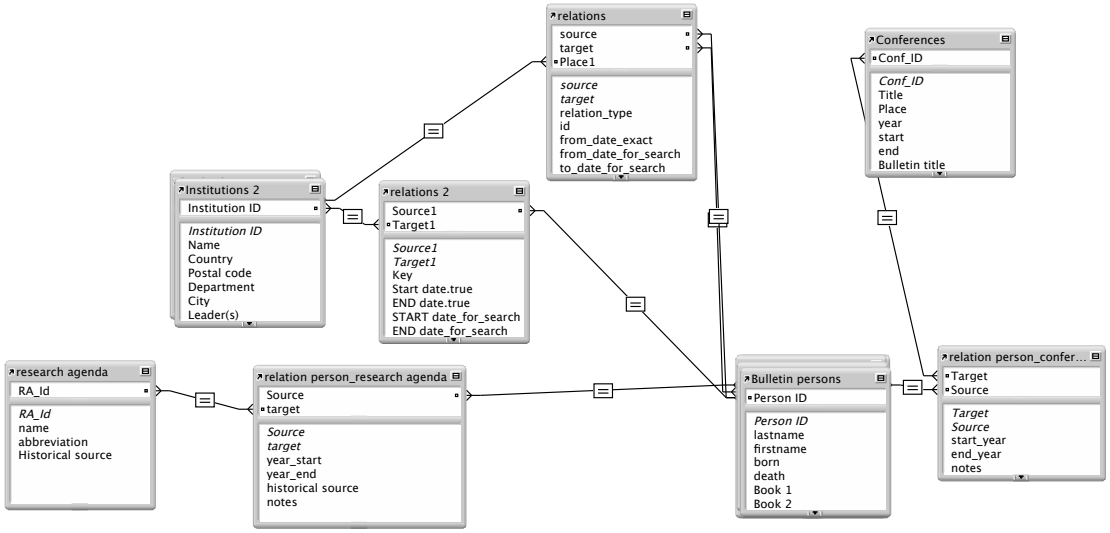


Abbildung 7.2: Relationale Struktur der Datenbank zur Geschichte der Allgemeinen Relativitätstheorie. Ausdruck der Strukturansicht von Filemaker 13

7.3.1 Die Genese der Datengrundlagen

Die Daten basieren auf der manuellen Analyse unterschiedlicher Quellen, wobei die wesentlichste das *Bulletin on general relativity and gravitation* [37] darstellt.[14] Alle Quellen sind in der Datenbank für alle wesentlichen Einträge dokumentiert. Abbildung 7.2 gibt einen Überblick über die relationale Struktur der Ausgangsdatenbank.

Sie umfasst 625 Personen und 328 Institutionen, wobei der Personendatensatz den Kern der Arbeit repräsentiert. Unterschiedliche Relationen beschreiben die internen Beziehungen zwischen Personen sowie der Personen zu Institutionen und Forschungsfeldern. Auf die Rolle der einzelnen Daten kommen wir im Folgenden noch zurück. Für eine Reihe von Personen sind hier bereits Verweise auf Normdaten vorhanden. Die vorhandene Struktur unterscheidet nicht konsequent zwischen Relationen zwischen Personen und anderen Entitäten auf der einen Seite und Attributen von Personen auf der anderen Seite. Das ist besonders problematisch für die Zuordnung zu Nationalitäten und Disziplinen. Nationalität und Disziplin sind Felder in der Tabelle der Personen. Es existieren Felder der Form *nationality_1* mit dazugehörigen Feldern *start_nationality 1* und *end_nationality 1*. Ähnliches gilt für die Disziplinen. Forschungsprogramme und Forschungsfelder sind jedoch in eigenen Tabellen beschrieben. Zusätzlich wurden an mehreren Stellen, wie z. B. im Falle der Nationalitäten, keine kontrollierten Vokabulare benutzt, so dass gleichbedeutende Einträge unterschiedlich verzeichnet sind. Die Datenbank ist ursprünglich als eine Form von elektronischem Notizbuch entstanden, ohne den Anspruch, einmal als Datengrundlage für systematische Auswertungen dienen zu sollen. Die dabei entstandenen Inkonsistenzen waren insoweit unproblematisch, als sich die Anzahl der Daten leicht überschauen lässt und die Datenbank damit ihre Aufgaben für die ursprüngliche Anwendung als Gedächtnisstütze erfüllt.

[14] Siehe dazu auch die detaillierte Darstellung der Quellen in [204].

7.3.2 Modellierung und Transformation

Die Überführung der Ausgangsdaten in eine Graphendarstellung folgt im wesentlichen den sich aus den Forschungsfragen ergebenden strukturellen Überlegungen. Zugleich sollen möglichst Synergien mit anderen Projekten genutzt werden. Insbesonders soll eine Verbindung der Daten mit den Daten der folgenden Fallstudie (Abschnitt 8) und die Anbindung an externe Datenbanken möglich sein.[15] Außerdem sollen vorrangig nur Strukturen modelliert werden, die auch durch Daten belegbar sind. Im Zentrum steht ein sehr einfaches Datenmodell, das im ersten Schritt die institutionellen und interpersonellen Beziehungen widerspiegelt.

Unsere Hauptklassen sind hierbei **E21_Person** und **sr:Institution**. Für die Personen haben wir uns entschieden, die entsprechende Klasse aus CRM beizubehalten. Für die Institutionen benutzen wir wie in allen anderen Fallbeispielen eine eigene Klasse, da wir in ihnen Zwischenkonstrukte von Akteuren und nominellen Gruppenbezeichnungen sehen. Sie sind daher Unterklassen sowohl von **E40_Legal_Body** als auch von **E78_Collection**. Im Hinblick auf eine historische Interpretation der entsprechenden Einrichtungen ist unsere Intention, dies zukünftig im Sinne der in den Einleitungskapiteln geschilderten Frage der Institutionalisierung weiter auszumodellieren. Letztendlich steht hinter der Frage, wann Institutionen sich als Akteure formieren, eine unserer Grundfragen an das *soziale Netzwerk*, womit sich auch die Hoffnung verbindet, dass im Laufe der weiteren Anreicherung des Modells logische Widersprüche in Bezug auf die Zuordnung von Instanzen zur Klasse **sr:Institution** Hinweise auf den Stellenwert dieser Instanzen geben können. In der praktischen Anwendung für die netzwerktheoretischen Untersuchungen schlägt sich die Interpretation der Rolle der Institutionen in der Gewichtung der durch sie erzeugten Bindungen nieder.

Die Relationen zwischen Personen teilen wir in zwei Hauptgruppen ein: eine Gruppe von symmetrischen, jedoch nicht notwendig transitiven Relationen, und eine Gruppe aus unsymmetrischen und ebenfalls nicht transitiven Beziehungen.[16] Die erste Gruppe spiegelt Kooperationsbeziehungen, die andere im weitesten Sinne Beziehungen der Einflussnahme bzw. umgekehrt des Ratsuchens wider.[17] Abbildung 7.3 zeigt diese Struktur. Beziehungen zwischen Personen und Institutionen teilen wir in vier tentative Gruppen ein, die in Abbildung 7.4 aufgezählt sind.

Alle Beziehungen besitzen in der Regel einen Zeitraum und im Falle der Personenbeziehungen einen Ort. Wir können die Beziehungen daher nicht als Objektrelationen auffassen, sondern verstehen sie im Sinne von CRM als Ereignisse,[18] die wir mit den entsprechenden Eigenschaften typisieren. Diese Typen bilden wir in einem einfachen Thesaurus ab. Später werden wir diese Typisie-

[15] Siehe Kapitel 6.8 zur Rolle von Linked Open Data.

[16] Die Fragen nach der Transitivität wird später mittels SIENA und ERGM noch zu untersuchen sein.

[17] Dies entspricht in etwa der von Ziberna und Lazega in [277] gemachten Klassifikation von Beziehungen in *advice-seeking* und *collaboration*.

[18] Da wir in der Regel keine genaueren Informationen über den Anfang oder das Ende von Aufenthalten an Institutionen oder auch die Beziehungen zwischen Personen haben, verzichten wir darauf, den Beginn und das Ende von Aufenthalten als Statusänderungen zu modellieren. Wir verstehen stattdessen den gesamten Aufenthalt als jeweils ein Ereignis, wobei es uns zukünftig offen steht, die Struktur zu erweitern und Anwesenheiten bzw. Beziehungen als **E28_Conceptual_Object** auffassen, das dann jeweils mit entsprechenden Ereignissen zu dessen Entstehung und Auflösung verbunden werden würde. Zum jetzigen Zeitpunkt würde ansonsten lediglich eine erhebliche Anzahl von Relationen ohne jegliche Information erzeugt werden.

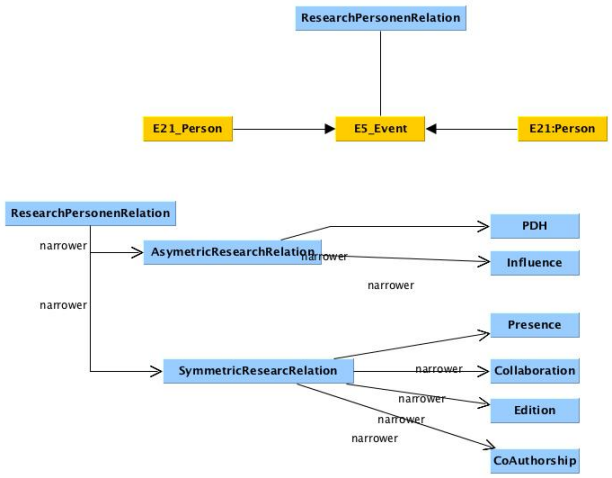


Abbildung 7.3: Personenbeziehungen und deren Klassifizierung.

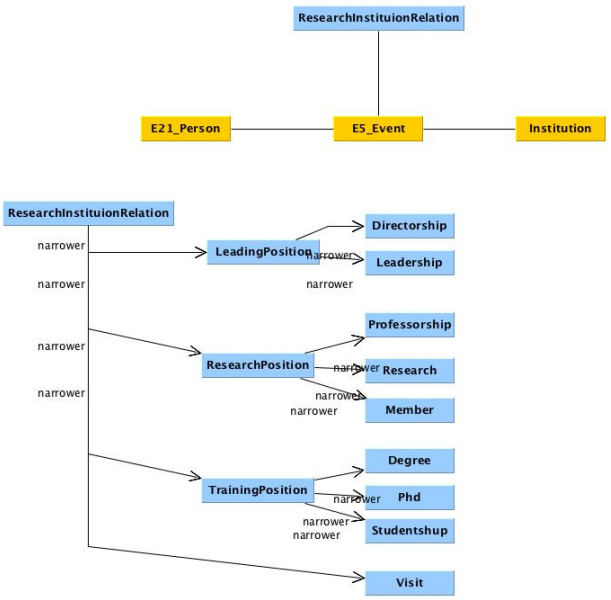


Abbildung 7.4: Beziehungen von Personen und Institutionen

rungen wieder nutzen, da wir damit Gewichtungen in den sozialen Netzwerken verbinden werden. Die Relationen zwischen Personen über die entsprechenden Events sind für den symmetrischen Fall **sr:collaborated** bzw. **sr:was_present_at** (oder Unterrelationen) sowie für den unsymmetrischen Fall von **sr:was_influencial** bzw. **sr:influenced_by**. Für die Beziehungen zwischen Personen und Institutionen benutzen wir Unterrelationen von **sr:person_institution_relation**. Analog bilden wir Klassen und Beziehungen für die Zugehörigkeit zu Disziplinen bzw. Forschungsfeldern. Diese bilden wir als **sr:ResearchPhase** ab, der über **sr:works_in_field** ein **sr:ResearchField** als Typ zugeordnet wird. Analog gehen wir mit der Zuordnung der Nationalität vor (Abbildung 7.5). Namen, ID der Datenbank, sowie GND- und VIAF-Identifier bilden wir jeweils als **sr:Identifier** ab. Die transformierten Daten legen wir in einem Graphen im *Triplestore* ab. Der Graph enthält 141.790 Tripel (Abbildung 7.6). [19]

7.4 Auswertungen und Interpretation

Von nun an dient die Datengrundlage im *Triplestore* als Ausgangspunkt für alle weiteren Auswertungen. Da dieses Fallbeispiel auch zur konkreten Diskussion über die Verfahren und zur Diskussion der Nachvollziehbarkeit der einzelnen Schritte in der Arbeitsgruppe dient, sind alle weiteren Schritte der

[19] Die Migration erfolgt nicht direkt aus Filemaker, sondern über einen Export in das xlxs-Format. Diese Daten werden im Repositorium unter *doi:21.11103/FK2.5XQP2X/* versioniert vorgehalten, so dass die weiteren Schritte nachvollziehbar bleiben. Danach erzeugen wir erst eine GraphMl-Datei aus den Daten mittels *relativity_analysis_0.14-generate-graph.ipynb*. Auch diese wird im Repositorium gesichert und aus dieser wird dann mit *From graph to RDF events.ipynb* die RDF-Darstellung erzeugt. Der Umweg über den Graphen ist ein Artefakt der ersten Arbeiten mit Daten und den ersten Experimenten mit der Graphdarstellung. Das Skript zur Umwandlung von Filemaker-Dateien über Excel in Graphen findet immer noch auch in anderen Kontexten seine Anwendung, um schnell einen Überblick über Netzwerkstrukturen zu gewinnen.

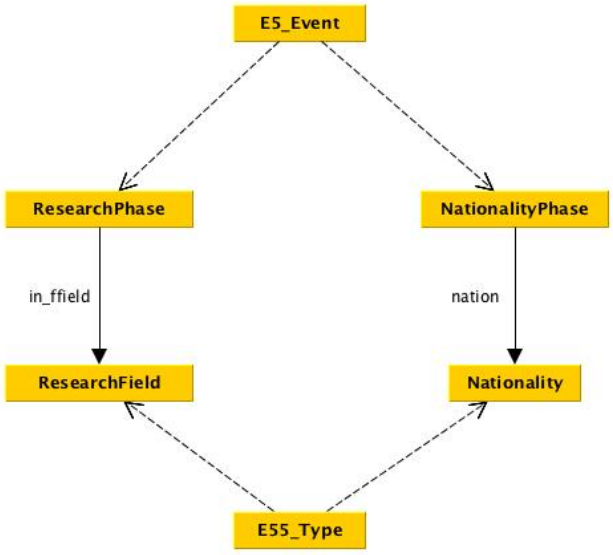


Abbildung 7.5: Nationalität und Forschungsfelder.

Analyse nun in Pythonnotizbüchern dokumentiert.

Zeitliche Entwicklung und die Rolle des Nachhallfaktors – ein erster Überblick.

Die Rolle des Nachhallfaktors, also des Parameters, der bestimmt, wie lange eine einmal aufgebaute Verbindung bestehen bleiben soll,[20] ist eine der grundlegenden Fragestellungen bei den zukünftigen Untersuchungen. In der ersten Phase des Projektes hatten wir zuerst für unterschiedliche Parameter unterschiedliche Entwicklungen konstruiert und dann qualitativ mittels Visualisierungen verglichen. Die Unterschiede zwischen den Graphen insbesondere auch im Vergleich zur Annahme, dass Verbindungen nicht aufgelöst werden, sondern sich im Gesamtgraphen als akkumuliertes Hintergrundwissen wiederfinden, bedürfen jedoch einer gründlicheren Analyse. Das erste Notizbuch[21] dient dazu, die Auswirkungen des Parameters auf die Netzwerkstruktur näher zu untersuchen. Wir lesen dazu die Daten aus dem Triplestore ein und erzeugen daraus den entsprechende Graphen. Dazu dienen Routinen aus dem SPARQL-Graph-Paket.[22] Wir erzeugen nun die Verlaufsgraphen für eine Reihe von Parametern für den Nachhall. Diese sind die Werte von 0 bis 10 sowie der Wert 100; letzterer entspricht der Akkumulation. Dieses wird einmal für den Fall berechnet, in dem nur die Personen-Personen-Beziehungen berücksichtigt werden, und in einem zweiten Fall unter Berücksichtigung des erweiterten Graphen, in dem wir Personen-Institutionen-Beziehungen auf Personen-Personen-Beziehungen projizieren.

Diese Ergebnisse vergleichen wir mit zufälligen Graphen mit den entsprechenden Parametern. Einmal nehmen wir gleiche Knoten- und Kantenzahl (Erdős-Rényi) und damit im Wesentlichen eine vollständig zufällige Strukturbildung an. Das andere Mal gehen wir von einer gleichen Knotenzahl

[20] Siehe die Abschnitte 8.5.1 und 7.2.

[21] *analyse_periods-first_overview _whole_graph_simulations.ipynb*.

[22] Siehe Abschnitt 6.2.3.

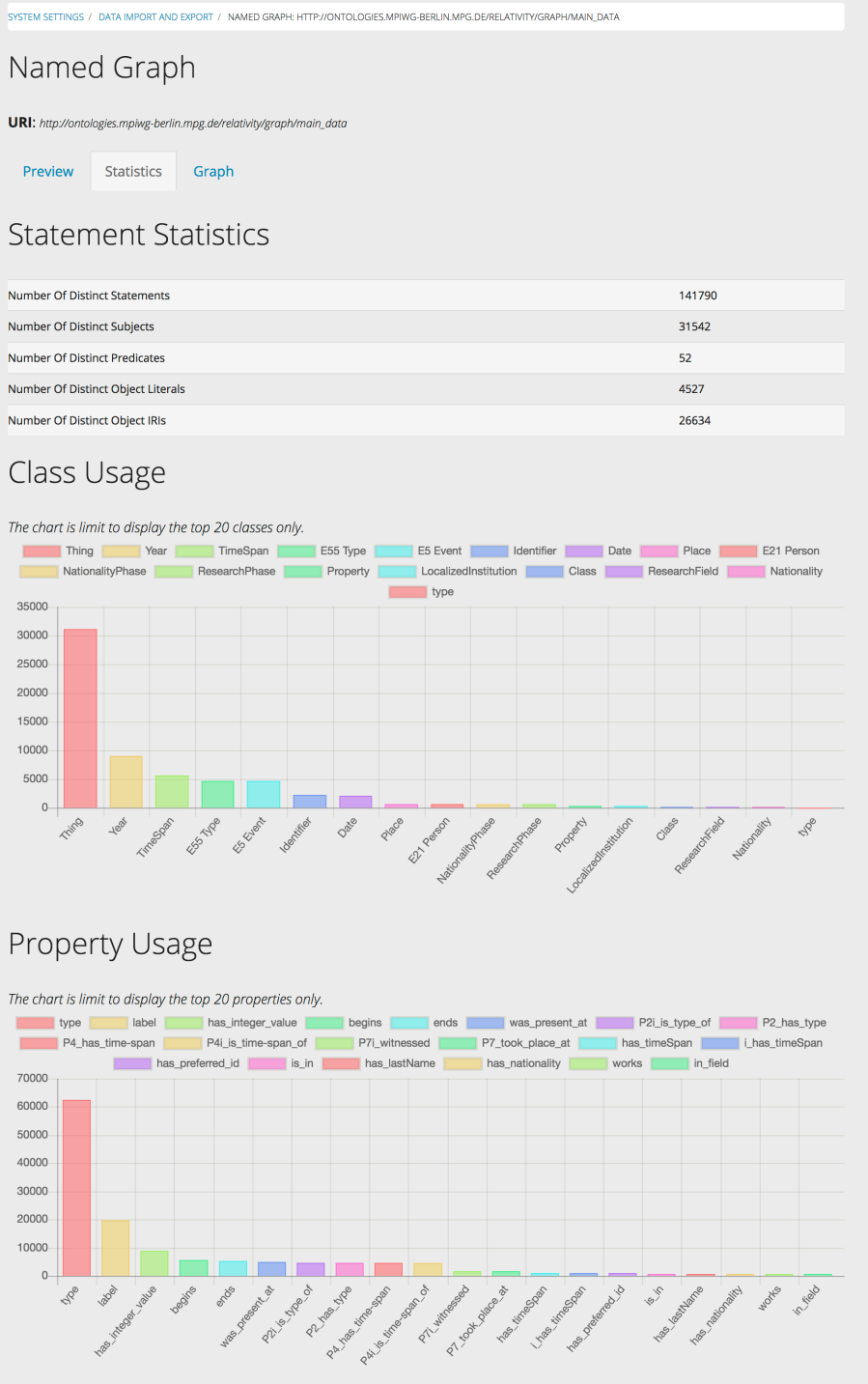


Abbildung 7.6: Übersicht über den Hauptgraphen des Projektes zur Allg. Relativitätstheorie

und gleichen Degree-Verteilung (Barabási) und damit ein skalenfreies Modell an.[23] Wir betrachten insbesondere die Entwicklung der größten Komponente in den unterschiedlichen Modellen. Bei der Modellierung der zufälligen Graphen nehmen wir an, dass es keine Abhängigkeiten zwischen den einzelnen Jahren gibt.[24] Daher werden für jeden Zeitabschnitt jeweils mehrere zufällige Graphen gebildet und wir arbeiten auf den Mittelwerten. Es ist außerdem möglich, die Anzahl der Iterationen zu ändern. In unserem Beispiel werden jeweils Mittelwerte aus 10 Graphen gewählt. Wir berechnen dann die quadratischen Abweichungen des realen Graphen von den gemessenen und normieren mit dem Mittelwert. Diese Werte sind dann ein Maß für den Abstand unserer realen Graphen von den Simulationen. Wir untersuchen des Weiteren die Entwicklung der größten Komponente als die für unsere Frage zunächst relevanteste Struktur im Hinblick auf die Formation des Forschungsfeldes der Allgemeinen Relativitätstheorie.

Die Ergebnisse (Abb. 7.7,7.8,7.9,7.10), zeigen zunächst sehr deutlich, dass unsere Graphen nicht das Ergebnis einer zufälligen Entwicklung sind. Sie zeigen jedoch auch, dass die Abweichungen nur gering von der Wahl der Länge der Zeitschnitte abhängen. Das Verhalten der Abweichungen von der Simulation ändert sich jedoch im Falle des einfachen Modells nach Erdős-Rényi (Abb. 7.7,7.8) deutlich, wenn institutionelle Bindungen im Netzwerk vernachlässigt werden.[25]

Die größte Abweichung von zufälligen Graphen erhalten wir, wenn wir $c = -3$ oder -2 Jahre annehmen. Ein Jahr ($c = -1$) nehmen wir aus inhaltlichen Überlegungen nicht in die Auswahl. Der Anstieg der einzelnen Parametern macht sich bereits bereits bei $c = -4$ Jahren bemerkbar. Für alle folgenden Untersuchungen nehmen wir daher im folgenden zum Vergleich jeweils die Parameter $-100, -8$, und -3 (-8, da er in etwa im mittleren Bereich der Entwicklung liegt) an und schauen uns hier zunächst die Unterschiede im Verlauf des gesamten Zeitraumes an.[26]

[23] Zu beiden Modellen siehe 5.5.

[24] In 7.6 gehen wir darauf noch einmal ausführlicher ein und betrachten dort auch Modelle, die eine zeitliche Entwicklung beinhalten.

[25] Die deutlichen Abweichungen konnten wir bisher noch nicht interpretieren, andererseits gehen unsere Annahmen davon aus, dass neue Bindungen vor allem in der zweiten Phase der Entwicklung des Feldes bewusst im Hinblick auf bereits etablierte Strukturen gewählt werden, so dass das Modell von Barabási und Albert den historischen Annahmen näher kommt.

[26] Die Standardabweichungen des Verhältnisses von Knoten und Kanten (grün) liegen im Bereich zwischen 0.03 und 0.06, während bei der Dichte Abweichungen zwischen 0.05 und 0.3 auftreten, daher benutzen wir erstere als Orientierung.

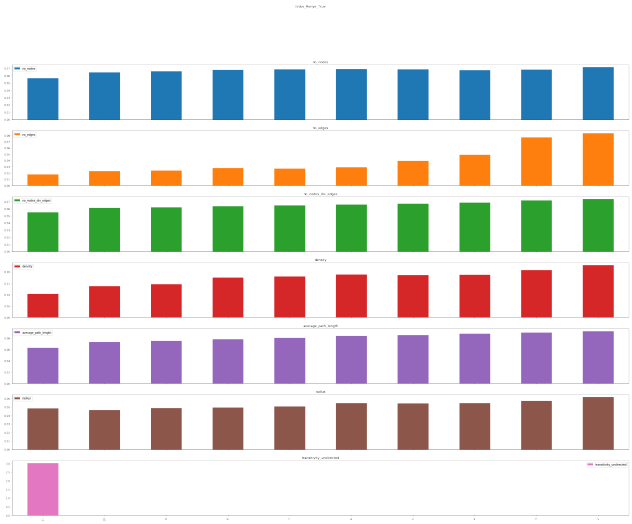


Abbildung 7.7: Vergleich einzelner Parameter der größten Komponente des realen Graphen und der Simulation eines zufälligen Graphen nach Erdős und Rényi, alle Beziehungen

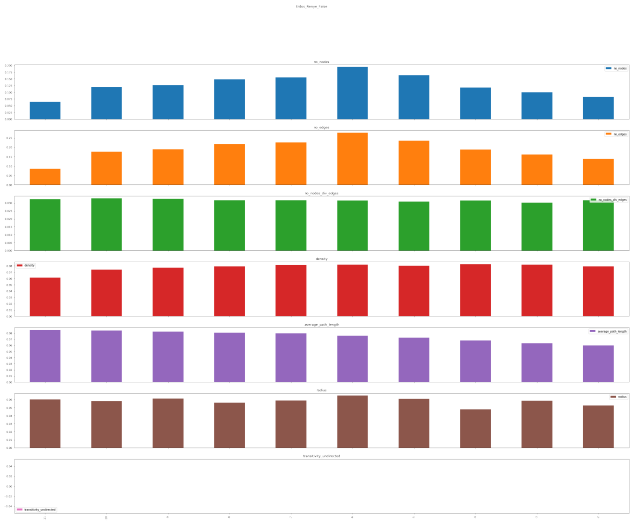


Abbildung 7.8: Vergleich einzelner Parameter des realen Graphen und der Simulation eines zufälligen Graphen nach Erdős und Rényi, nur direkte Personenbeziehungen

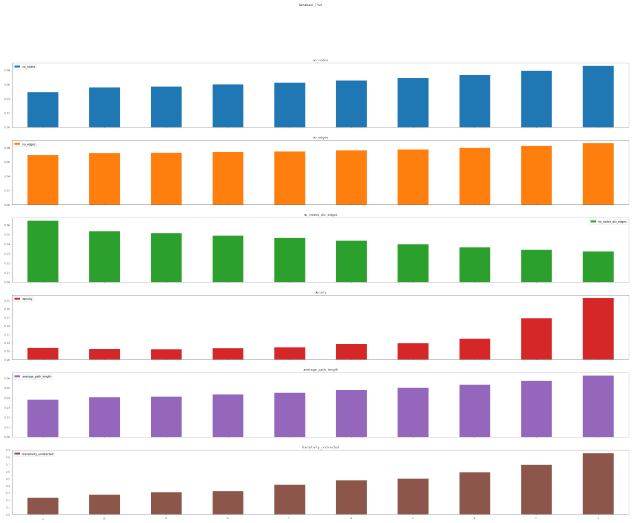


Abbildung 7.9: Vergleich einzelner Parameter der größten Komponente des realen Graphen und der Simulation eines zufälligen Graphen nach Barabasi, alle Beziehungen

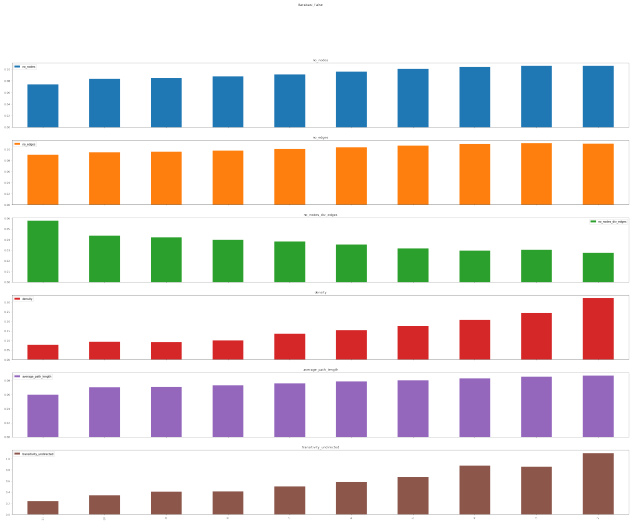


Abbildung 7.10: Vergleich einzelner Parameter der größten Komponente des realen Graphen und der Simulation eines zufälligen Graphen nach Barabasi, nur direkte Personenbeziehungen

7.4.1 Entwicklung der Personennetzwerke - Details

Nach dem ersten Überblick gehen wir der Frage nach, wie sich unsere Netzwerke über die Zeit entwickeln. Auch hier interessiert uns zunächst der Einfluss des Nachhallparameters auf die strukturellen Verläufe. Für die Experimente steht ein Pythonnotizbuch[27] zur Verfügung. Dieses erlaubt die Veränderung des Nachhallparameters und erstellt darauf basierend Visualisierungen über die Zeitverläufe. Die Abbildungen 7.11,7.12,7.13,7.13 zeigen die Entwicklungen der größten Komponente des Netzwerkes über den Zeitverlauf mit dem Nachhallparameter -3 für das erweiterte Netzwerk und das Netzwerk ohne die institutionellen Beziehungen.

Die Diagramme bestätigen die Vermutung zweier getrennter Entwicklungen vor und nach dem Zweiten Weltkrieg. Es zeigt sich außerdem ein großer Umbruch nach 1951, und auch der vermutete Anstieg mit dem Höhepunkt in den 1960er Jahren wird deutlicher. Letzterer wird verstärkt, wenn wir nicht den erweiterten Graphen betrachten und nur auf die explizit dokumentierten Personenbeziehungen zurückgreifen, ohne mögliche institutionelle Beziehungen in Betracht zu ziehen. In diesen Fällen wird eine Plateauphase nach einem sehr schnellen Anstieg zu Beginn der 1950er Jahre bereits in der Mitte der 1950er erreicht. Diese Abweichungen lassen mehrere Schlüsse zu, die von der Bewertung der durch die Institutionen möglicherweise verborgenen Beziehungen auf der einen und der Bewertung der Renaissance der ART auf der anderen Seite abhängen: Entweder sind versteckte institutionelle Beziehungen nicht oder nur wenig für die Entwicklung des Feldes relevant oder die dokumentierten Beziehungen überhöhen die Rolle der 1960er Jahre. Der Verlauf der Kurven in den ersten beiden Jahren 1930/1931 ist ein Problem unserer Datenlage zu Beginn unseres Zeitraumes mit nur wenigen Knoten vor 1931. In der akkumulierten Darstellung führt dies am Anfang zu einer statistischen Abweichung, die mit der Größe des Nachhallparameters wächst. Für -3 können wir jedoch davon ausgehen, dass die Ausschläge zumindest ab 1934 auch in der absoluten Höhe weitestgehend gesichert sind. Um so auffälliger ist hier, dass die relative Größe der größten Komponente vor dem Krieg in diesem Falle erst wieder in den 1960er Jahren erreicht wird. Eine weitere stärker versteckte Entwicklung im Verlauf der Graphen wird deutlich, wenn wir als Vergleichsgröße wieder die zufälligen Graphen heranziehen. Auch dieses wird mit dem Notizbuch ermöglicht. Die Abbildungen 7.15, 7.16 und 7.17 zeigen das Ergebnis für den Fall der 3 Jahre wieder in beiden Situationen – für den erweiterten und den einfachen Graphen im Vergleich mit der Simulation. Insbesondere 7.17 macht hier eine Entwicklung deutlich, die wir in der historischen Interpretation vermutet hatten, die aber in den Diagrammen, die lediglich die Entwicklung des Netzwerkes über die Zeit darstellen, nicht signifikant ist.

[27] *Dynamische Entwicklung der Graphen - im Vergleich zu zufälligen Graphen.pynb*

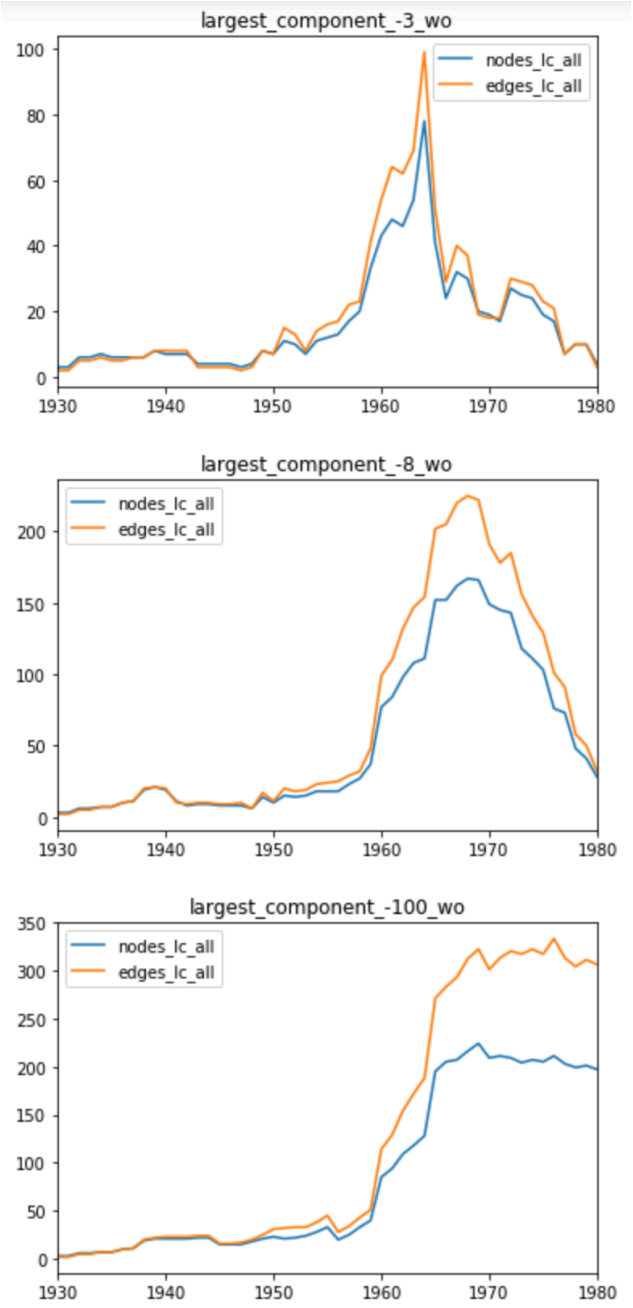


Abbildung 7.11: Knoten und Kanten der größten Komponente

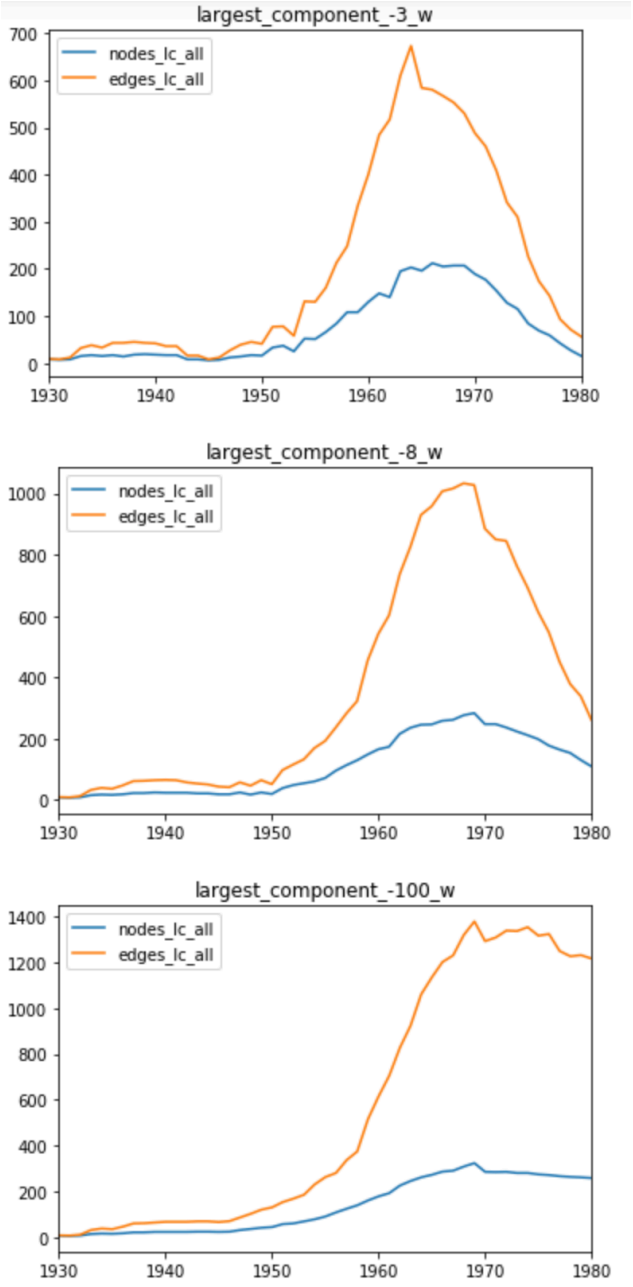


Abbildung 7.12: Knoten und Kanten der größten Komponente - erweitertes Netzwerk

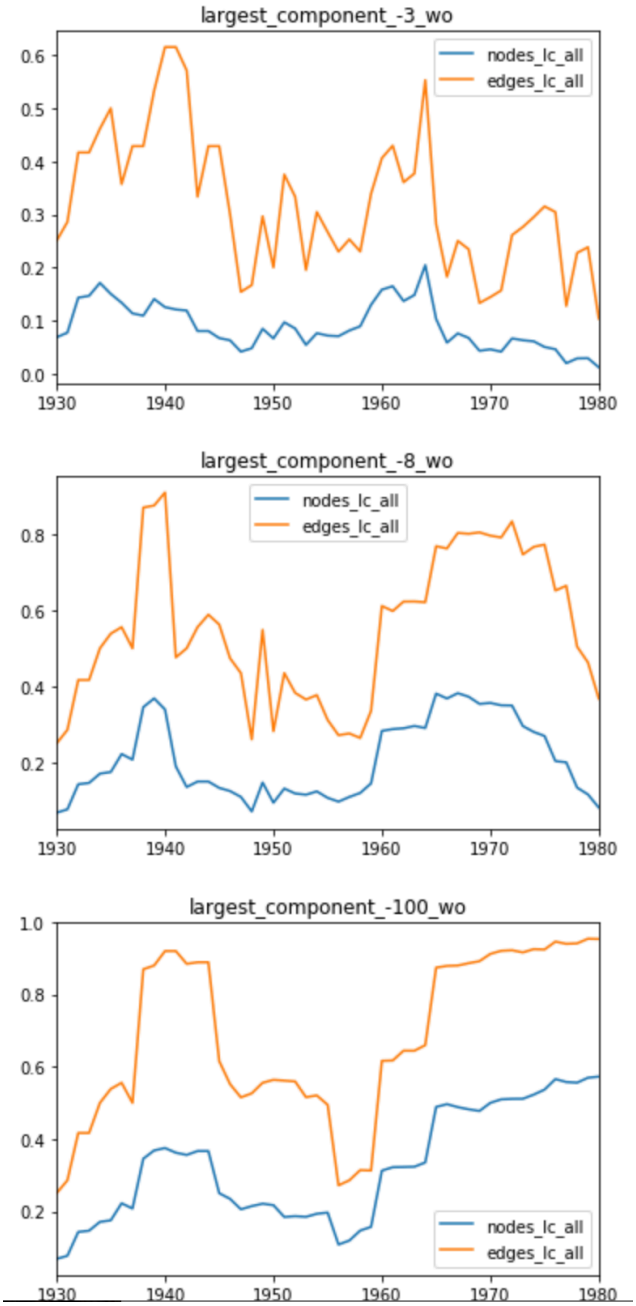


Abbildung 7.13: Knoten und Kanten der größten Komponente, Verhältnis zu allen Kanten und Knoten

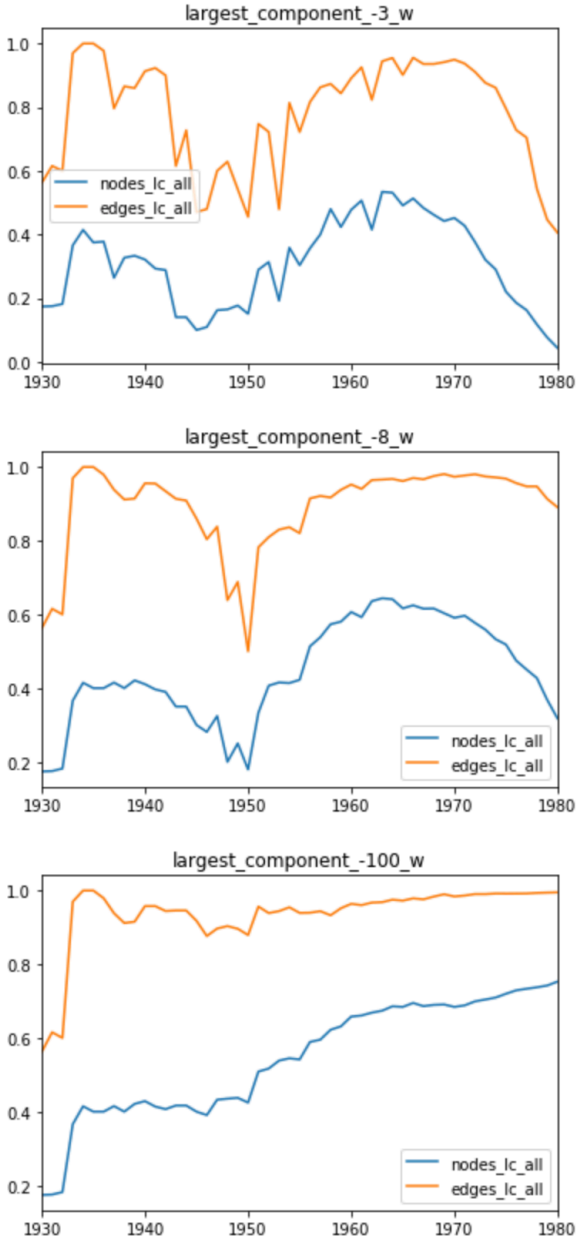


Abbildung 7.14: Knoten und Kanten der größten Komponente, Verhältnis zu allen Kanten und Knoten - erweitertes Netzwerk

Wir beobachten hier ein deutlich verspätetes Einsetzen der Bildung der ersten größeren Komponente im Vergleich zur Simulation; der Einbruch beginnt bereits vor den 1940er Jahren. Im Jahre 1935/1936 fällt die Entwicklung des realen Graphen bereits hinter den simulierten Graphen zurück. Dafür ist der Abfall in den Kriegsjahren deutlich weniger stark als eine zufällige Entwicklung erwarten ließe, der große Abfall folgt erst in den ersten Jahren nach Kriegsende. Diese Beobachtung gilt für beide Formen der Simulation. Für den akkumulierten Graphen (-100) ist der Einbruch ab etwa 1930 noch einmal größer. Der reale Graph bleibt danach immer hinter dem simulierten zurück. Das

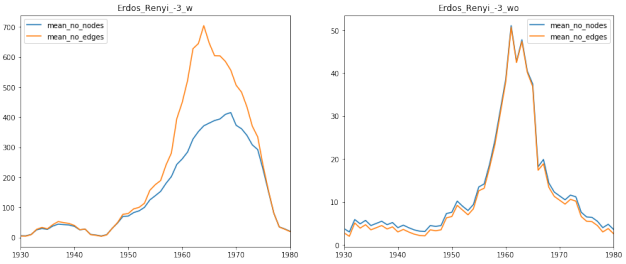


Abbildung 7.15: Verlauf der Entwicklung der größten Komponente für simulierte Netzwerke, 3 Jahresschnitte, ohne “met_at“, Erdos-Renyi

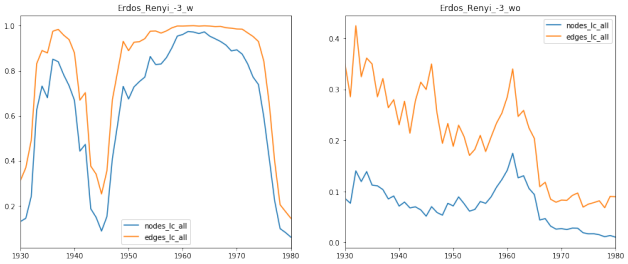


Abbildung 7.16: Verlauf der relativen Entwicklung der größten Komponente für simulierte Netzwerke, 3 Jahresschnitte, ohne “met_at“, Erdos-Renyi

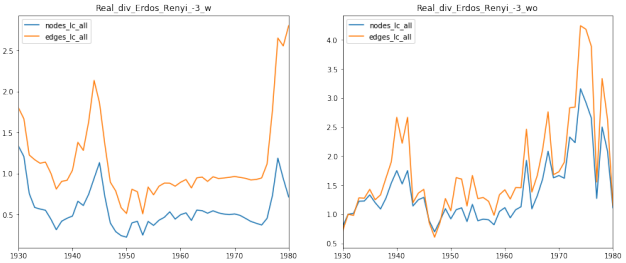


Abbildung 7.17: Verlauf der realen zur simulierten Entwicklung der größten Komponente für simulierte Netzwerke, 3 Jahresschnitte, ohne “met_at“, Erdos-Renyi

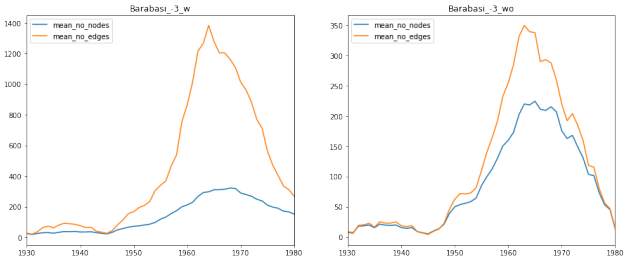


Abbildung 7.18: Verlauf der Entwicklung der größten Komponente für simulierte Netzwerke, 3 Jahresschnitte, ohne “met_at“, Barabasi

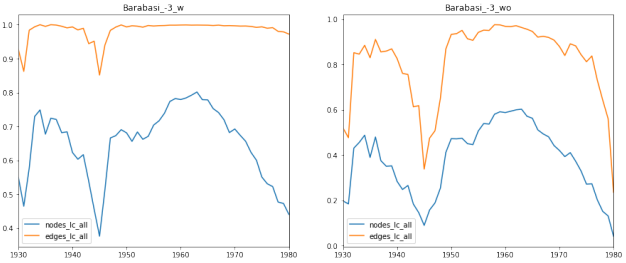


Abbildung 7.19: Verlauf der relativen Entwicklung der größten Komponente für simulierte Netzwerke, 3 Jahresschnitte, ohne “met_at“, Barabasi

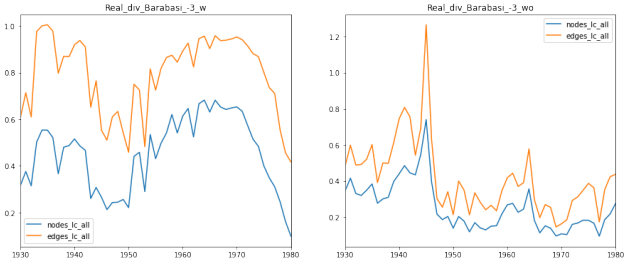


Abbildung 7.20: Verhält der realen zur simulierten Entwicklung für simulierte Netzwerke, 3 Jahresschnitte, ohne “met_at“, Barabasi

Ergebnis ist nicht überraschend, ist aber zumindest ein Indiz für die Qualität des Datensatzes und die vorgeschlagenen Methodik. Außerdem sehen wir im Jahr 1951 nicht nur in den absoluten Zahlen, sondern auch relativ zu den simulierten Netzwerken eine Kehrtwende: Ab dort beginnt der Graph wieder deutlich stärker zu wachsen und erreicht schnell wieder das Vorkriegsniveau. Die Beobachtungen sprechen für zumindest drei Phasen: eine bis 1936, ein weitere von 1936 bis 1951, sowie die Zeit ab 1951, wobei die Umbruchphase beginnend 1951 besonders deutlich ist. Der Vergleich der beiden Netzwerktypen – (w) mit allen Kanten und (wo) ohne die institutionellen Beziehungen – vergrößert die Effekte im Zeitintervall vor Beginn des Zweiten Weltkrieges bis in die 60er Jahre weiter. Im Fall (wo) ist das reale Ecken- bzw. Kantenverhältnis zwar immer kleiner als das simulierte, doch auch hier sind die Einbrüche verspätet und die Erholung ist noch deutlicher verzögert. Außerdem zeigt sich, dass das Modell nach Erdős und Rényi in der Nachkriegsphase näher am realen Graphen liegt als das Modell nach Barabási-Albert. In der Nachkriegsphase scheint die zufällige Wahl eine Wahl im Sinne des *preferential attachment*[28] zu überwiegen.

7.4.2 Time Slices – Erstes Fazit

Eine Variation der Länge des Nachallparameters ergibt das folgende Bild (7.21) für den Fall des erweiterten Netzwerkes. Je größer der Nachhallfaktor wird, desto besser stimmt die Simulation mit dem Modell nach Barabási (also mit *preferential attachment*) zumindest für die Anzahl der Knoten der größten Komponente mit den empirischen Daten überein. Die Entwicklung der Kantenanzahl bleibt hinter der Simulation zurück. Diese gilt für nahezu alle Zeiträume. Ähnliches gilt für Simulationen nach Erdos-Renyi, so dass in diesem Fall keine Einschätzung vorgenommen werden kann, welches Modell näher an den realistischen Daten liegt.

Betrachten wir die institutionellen Bindungen nicht, dann erhalten wir ein anderes Bild. In der Periode vor 1955 weicht Erdős-Rényi deutlich von den realen Daten ab, während nach 1960 sich die simulierten den realen Daten mit steigendem Nachhallfaktor annähern. Der Vergleich mit dem anderen Modell ist nicht eindeutig. Bei niedrigen Werten liegt die Simulation vor 1945 besser an den gemessen Daten als bei Erdős-Rényi. Im Zeitraum nach 1950 weichen sie bei kleinen Werten für den Nachhallfaktor deutlich von den realen Werten ab, erst mit deutlich steigendem Nachhallfaktor erhalten wir eine Annäherung. Der Abfall vor 1935 ist wahrscheinlich ein Artefakt der Daten, die erst 1930 beginnen.[29]

7.4.3 Rolle der Gewichtungen

Auf die Rolle der Gewichte sind wir bisher noch nicht näher eingegangen. Hier erlaubt ein Pythonnotizbuch[30] die Untersuchung, wie sich die Modularität[31] des Netzwerkes bezogen auf Eigenschaften wie Nationalität und Disziplinenzugehörigkeit über die Zeit verändert. Hier sind noch weitere Stu-

[28] Siehe Abschnitt 5.5.3.

[29] Andere Parameter, wie Dichte und Radius der größten Komponente verhalten sich ähnlich, dies kann mit den Notizbuch nachgeprüft werden.

[30] *Strukturelle Entwicklungen der Graphen - Disziplinen und Nation.pynb*

[31] Siehe 5.4.3

Abbildung 7.21: Vergleich zwischen realen und simulierten Graphen für verschieden Nachhallparameter, sowie mit und ohne institutionelle Bindungen

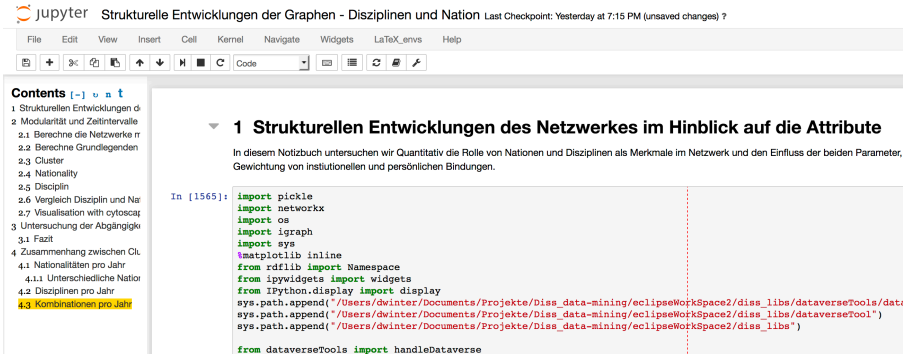


Abbildung 7.22: Strukturelle Entwicklungen der Graphen – Disziplinen und Nationen

dien notwendig. Insbesondere benötigen wir eine detailliertere Aufschlüsselung der Zuordnung der handelnden Personen sowohl in ihrer Selbsteinschätzung als auch aus der Außensicht durch eine genauere Analyse der Forschungsfelder. Die Datenlage ist hier bisher nicht für zuverlässige quantitative Auswertungen ausreichend. Es zeigen sich jedoch zumindest Tendenzen. Abbildung 7.23 zeigt für die beiden Perioden, in denen wir jeweils eine ausgeprägte große Komponente vorfinden, – wie grundsätzlich erwartet – dass die Modularität der Disziplinen stärker mit dem Ansteigen der Gewichte für die institutionellen Beziehungen abnimmt als die der Nationalitäten. Auffällig ist auch die Entwicklung, dass jeweils nach der Etablierung der größten Komponente beide Kurven abfallen. Wiederum ist das Jahr 1951 ausgezeichnet. Hier ist jeweils der Unterschied zwischen den Modularitäten am deutlichsten ausgeprägt, da sich die beiden Minima der Kurven in dieser Phase jeweils genau um ein Jahr verschieben. Zugleich ist in den 1960er Jahren – also der Hochphase der Renaissance der ART – die nationale Komponente deutlich stärker ausgeprägt. Erhöhen wir die Gewichtung der institutionellen Bindungen, so nehmen Disziplinarität als auch Nationalität als Faktoren für Modularität ab – der Faktor Disziplinarität jedoch geringer als der Faktor Nationalität. Die genaue Interpretation ist hier noch zu leisten. Eine erste Deutung scheint die Erwartung zu bestätigen, dass Institutionen in der späteren Phase multinationaler und multidisziplinärer ausgestaltet waren, als es die dokumentierten Einzelbeziehungen zeigen, wobei es eine leichte Tendenz in Richtung der Nationalisierung zu geben scheint.[32] Die genauere Interpretation und Prüfung im Einzelfall stehen hierbei aus. Die beiden Abbildungen 7.25 und 7.24 zeigen diese Entwicklungen noch einmal als Parameter des Gewichtsfaktors bei festem Nachhallfaktor $c = -8$.

7.4.4 Internationale und interdisziplinäre Kooperationen

Experimentiell sind wir in diesem Kontext noch der Frage nachgegangen, ob sich Korrelationen zwischen den Ergebnissen von Standardclustermethoden[33] und den Kriterien Nationalität und Disziplin ergeben. Das Notizbuch erlaubt die Untersuchung wieder unter Variation unserer freien Parameter.

[32] Bei diesen einfachen Betrachtungen benutzen wir die formalen Ausbildungsgänge zur Charakterisierung der disziplinären Zugehörigkeit. Wir sind uns natürlich bewusst, dass dies nur ein schwaches Merkmal für die tatsächlichen Arbeitsfelder ist. Es ist aber dennoch zunächst einmal ein formales Merkmal der Akteure. Die leichte Abnahme der Internationalität beinhaltet sicher auch noch Nachwirkungen der Migrationsbewegungen in der ersten Hälfte des 20. Jahrhunderts.

[33] Siehe 5.4.3.

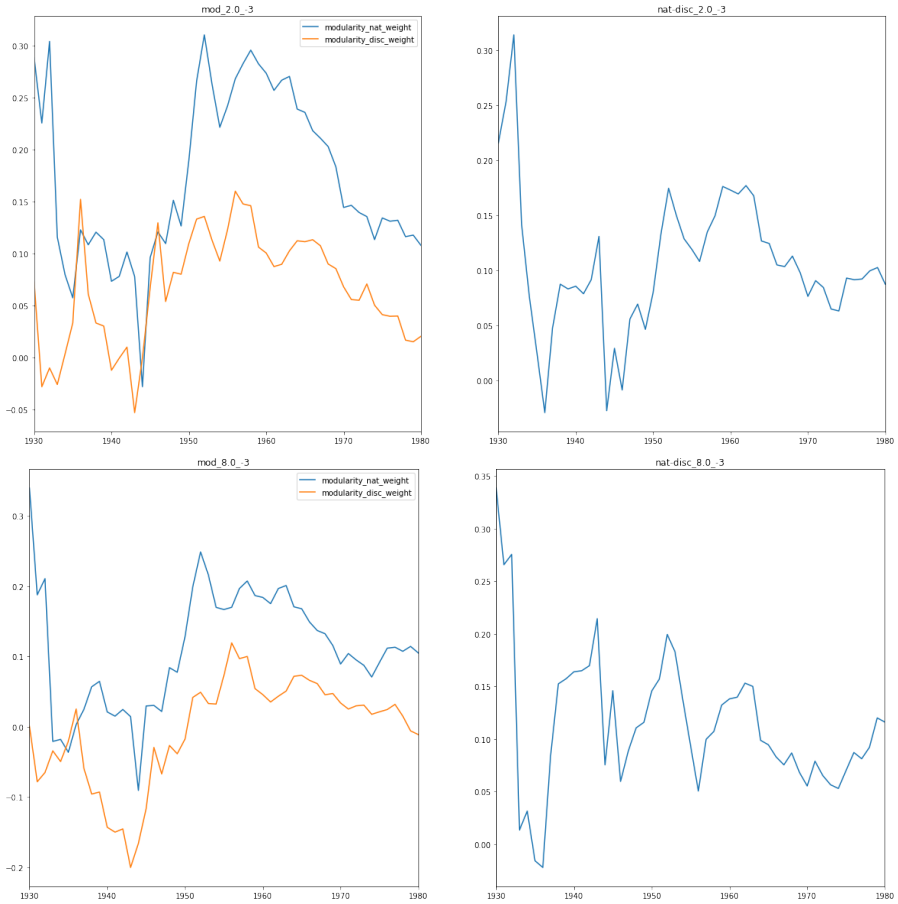


Abbildung 7.23: Veränderung der Modulariät von Disziplinen und Nationalität – Links: die absoluten Zahlen, rechts die Differenz von Nationalität und Disziplin, oben für den Gewichtfaktor 2 unten für den Faktor 8

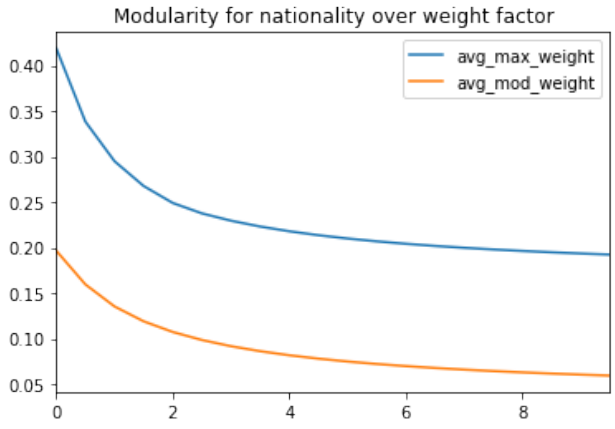


Abbildung 7.24: Entwicklung der Mittel- und Maximalwerte für Modularität der größten Komponente für das Kriterium Ṅationalitätüber dem Gewichtsfaktor mit Nachhallfaktor -8

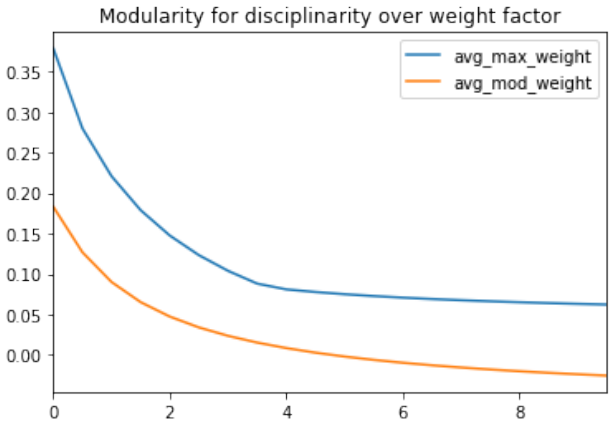


Abbildung 7.25: Entwicklung der Mittel- und Maximalwerte für die Modularität der größten Komponente für das Kriterium Disziplinüber dem Gewichtsfaktor mit Nachhallfaktor -8

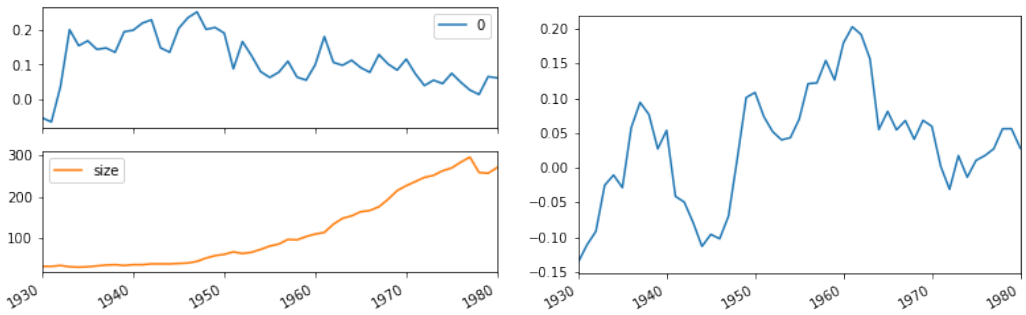


Abbildung 7.26: Korrelationen für Nachhallfaktor -8 und Gewicht 3,Cluster nach Infomap

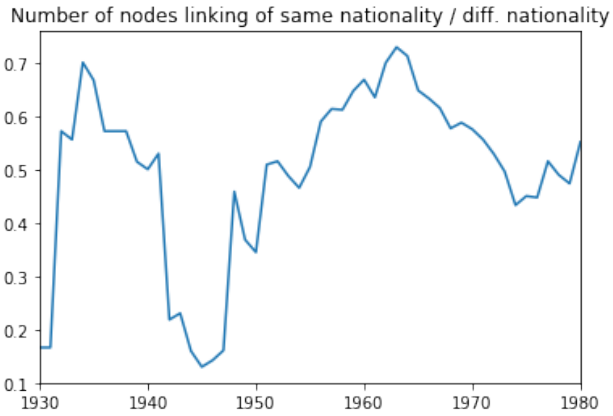


Abbildung 7.27: Anzahl der Knoten die Personen mit gleicher Nationalität verbinden relativ zu ungleicher mit Nachhallfaktor -8

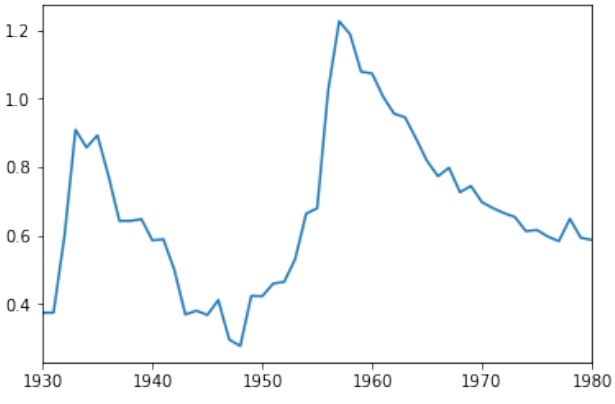


Abbildung 7.28: Anzahl der Knoten die Personen mit gleicher Disziplin verbinden relativ zu ungleicher mit Nachhallfaktor -8

Das Ergebnis entspricht für fast alle durchgespielten Kombinationen zumindest qualitativ dem in Abbildung 7.26 gezeigten Verlauf. Im Hinblick auf Spearman-Korrelation erhalten wir für die Disziplin eine Schwankung im statistischen Bereich um die Null-Lage und für die Nationalität die Tendenz einer auf Null abfallenden Korrelation bei gleichzeitig starkem Anwachsen der Anzahl der Cluster.

Das Notizbuch erlaubt zusätzlich die statistische Auswertung der Eigenschaften von Dyaden innerhalb des Knotens im Hinblick auf ihre Attribute (Abbildungen 7.27,7.28). Ziel ist es, eine Vorstellung von der Entwicklung nationaler und internationaler Kooperationszusammenhänge sowie deren disziplinärer Struktur zu gewinnen. Auch hier zeigt sich wieder eine Tendenz, dass sich in den entsprechenden zeitlichen Phasen (2. Weltkrieg, 1950/1951 und 1960er Jahre) signifikante Veränderungen der Struktur vermuten lassen, deren historische Interpretation jedoch noch aussteht, da die methodischen Probleme bei der Beurteilung der Bedeutung und Zuweisung von Nationalität und Disziplin bisher noch nicht hinreichend geklärt sind.

7.4.5 Personen und ihre Rollen

Neben den noch folgenden Untersuchungen zu Kozitationsnetzwerken liegt ein Hauptaugenmerk auf den Entwicklungen und Entwicklungstendenzen im Hinblick auf die zentralen Personen. Da sich die Rollen von Einzelpersonen in unseren Netzwerken mit einer relativ geringen Anzahl von Kanten durch veränderte Annahmen über die Gewichtung von Beziehungen und sowie durch das Wegfallen von nur wenigen Beziehungen deutlich verändern können, ist das Notizbuch *Personen - Entwicklung.ipynb* darauf ausgerichtet, die Entwicklungen von Personen über die Zeit nachvollziehen zu können. Insbesondere hängt die Bedeutung von Personen mit hoher Betweenness-Zentralität, die unterschiedliche – häufig über Nationalität – charakterisierbare Teilnetze verbinden, erheblich von Annahmen über den Beginn und das Ende dieser Beziehungen ab. Wir haben hier insofern ein typisches Small-World-Problem[34] in den Teilnetzen. Dieses gilt insbesondere für Phasen starker Separation zwischen den einzelnen Blöcken, wie etwa in den Zeiten des Kalten Krieges, wo wir die Quellenlage über Kontakte zwischen Personen auf beiden Seiten des Eisernen Vorhanges zumindest kritisch beurteilen müssen.[35]

Ohne Eingriffe in das Skript können Nachhallparameter und Gewichte verändert und Auswirkungen auf die Struktur des Netzwerkes untersucht werden. Als Beispiele zeigen die Abbildungen 7.29 und 7.30 den Verlauf von normalisierter Betweenness- und Closeness-Zentralitäten: Jeweils oben mit einem relativen Gewicht von Bindungen über die Institutionen von 2.3 und unten ohne Berücksichtigung von institutionellen Bindungen.

In der normierten Betweenness sehen wir grob zwei unterschiedliche Entwicklungen: eine in der Zeit vor etwa 1966, die andere danach. Mit Berücksichtigung institutioneller Bindungen sehen wir eine Tendenz zur Dezentralisierung in der ersten Phase, danach wiederum einen Anstieg. Ohne Berücksichtigung der institutionellen Bindungen ändert sich die Struktur ebenfalls etwa in der Mitte der 1960er Jahre. Zwar liegt jetzt die maximale Zentralität auf etwa gleichem Niveau ab den 1932 Jahren, ab Mitte der 1960er Jahre sind jedoch deutlich mehr Personen mit höheren Zentralitäten im Netzwerk vorhanden. Mit der interaktiven Komponente des Notizbuches können wesentliche Parameter des Netzwerkes direkt geändert und so die Auswirkungen auf die zeitlichen Entwicklungen untersucht werden. Das Notizbuch lässt den unmittelbaren Vergleich unterschiedlicher Parametrisierungen zu. So zeigen sich auffällige Veränderungen beispielsweise in der relativen Stellung von Wheeler, der bei $+10$ den deutlichen Spitzenplatz einnimmt, bei -10 jedoch nur noch mit geringen Wertungen auftritt. (7.31).

[34] Siehe Abschnitt 5.5.4.

[35] Hier macht sich wiederum das bereits in Abschnitt 7.1 erwähnte Problem der Quellenlage für den nicht anglo-phonen Bereich bemerkbar.

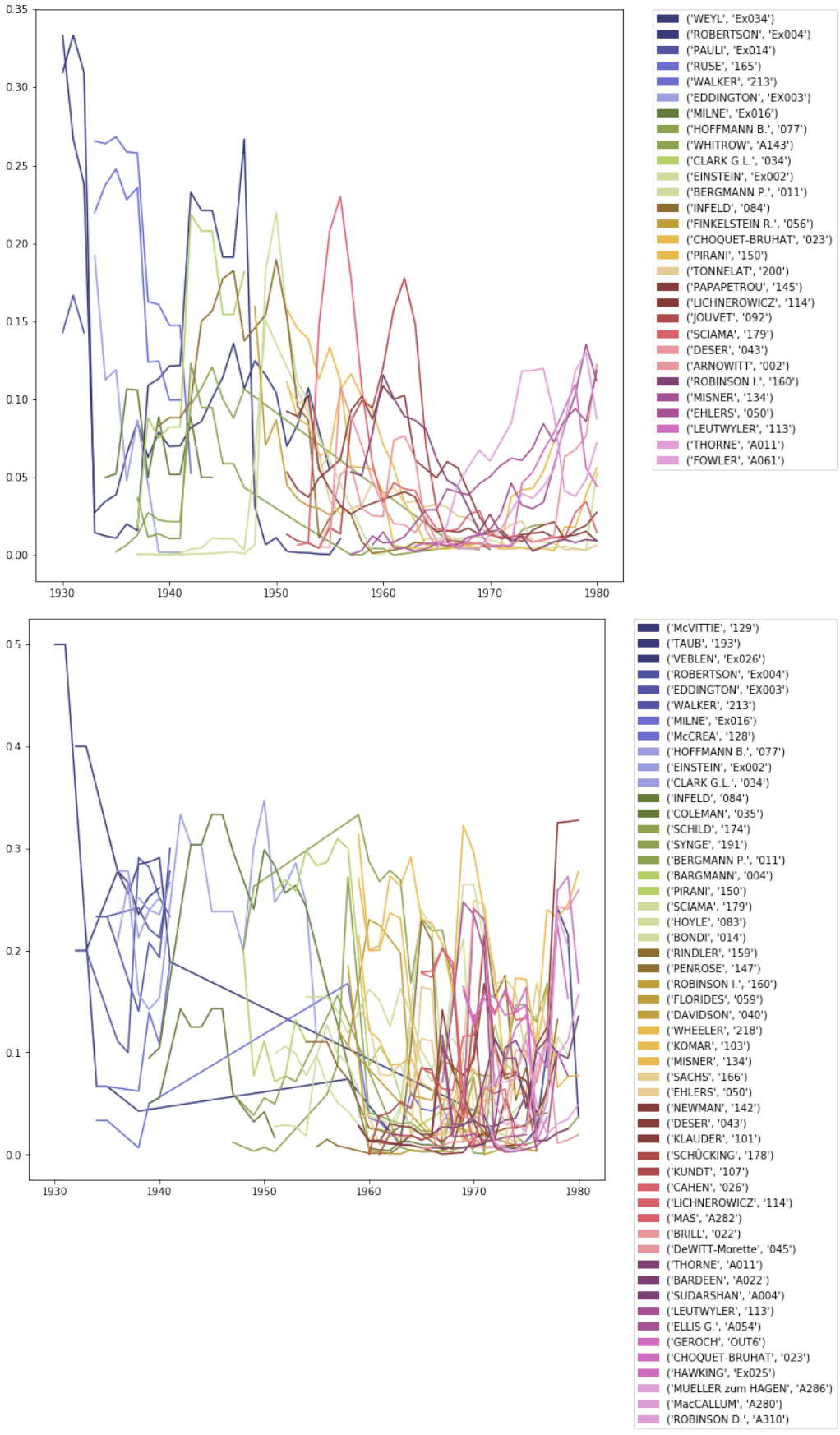


Abbildung 7.29: Illustration: Verlauf der Betweenness mit Nachhallfaktor -8 (größte Komponente). Personen, deren maximale normierte Betweenness über 0.1 liegt. Einmal mit Gewicht 2.3 einmal mit 0.

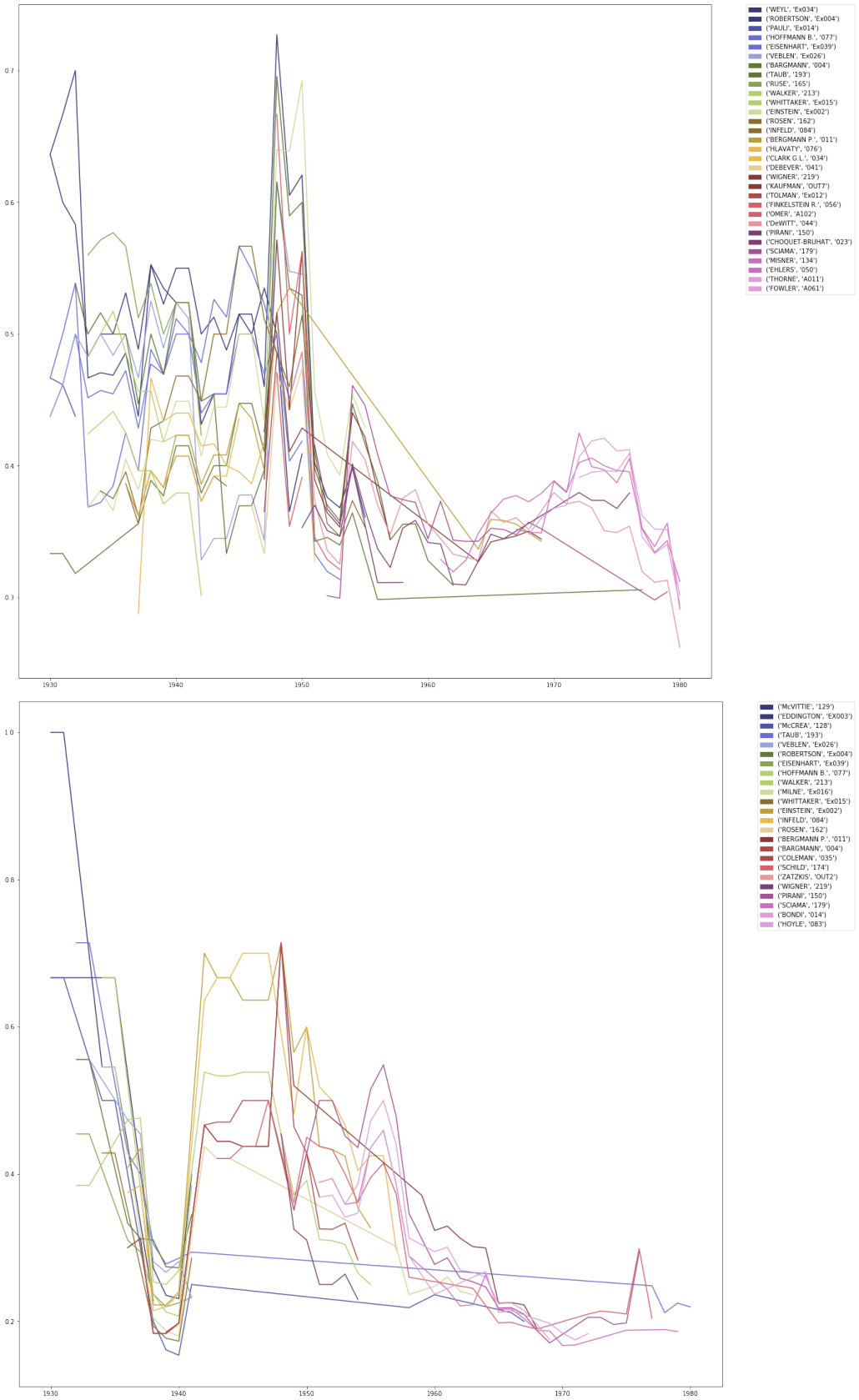


Abbildung 7.30: Illustration: Verlauf Closeness mit Nachhallfaktor -8 (größte Komponente), Personen, deren maximale Closeness über 0.4 liegt.

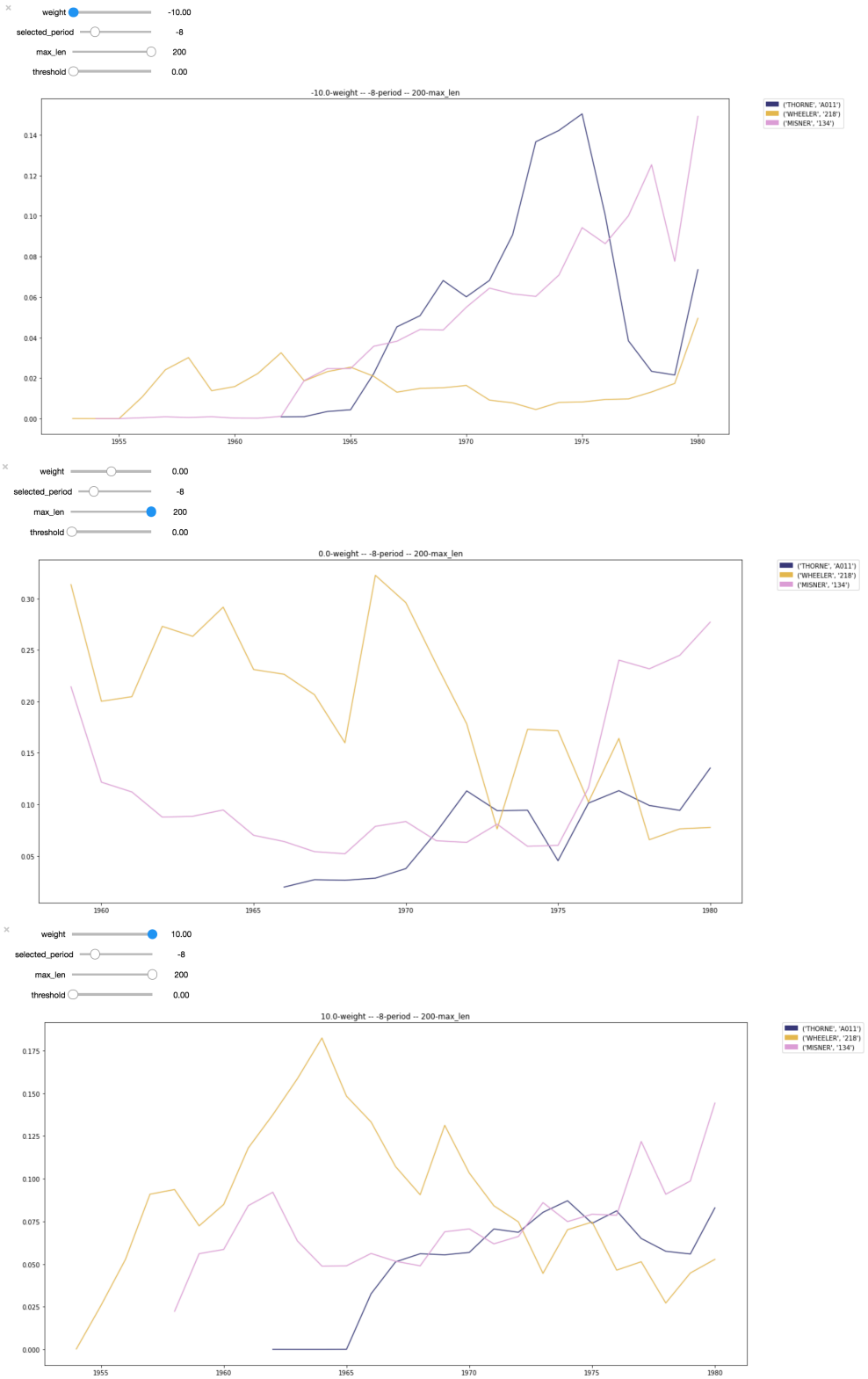


Abbildung 7.31: Ausschnitt aus dem Notizbuch: Vergleich unterschiedlicher Gewichtungen für Wheeler. -10 steht hier für k =1/10 und 0 für das Netzwerk ohne die institutionelle Erweiterung

7.5 Forschungsprogramme

Abschließend soll an dieser Stelle noch ein kurzer Blick auf die Rolle von Forschungsprogrammen erfolgen. In den vorhergehenden Betrachtungen haben wir Interdisziplinarität lediglich vor dem Hintergrund der sehr groben Einteilung von Bildungsgängen behandelt. Dies hat zunächst den Vorteil einer doch mehr oder weniger objektiven Zuordnung. Welche Arbeitsfelder von den einzelnen Wissenschaftlerinnen und Wissenschaftlern wirklich abgedeckt werden, ist damit sicher nur sehr eingeschränkt bestimmbar. Eine genauere Beschreibung ergibt sich durch die Zuordnung zu Forschungsfeldern. Wie schon in Abschnitt 7.3.2 beschrieben, sind diese Forschungsphasen, anders als die disziplinäre Zuordnung, zeitabhängige Größen. Ein vollständiges Modell muss die Formation und das Wechselspiel der Forschungsprogramme mit den anderen Komponenten des Netzwerkes darstellen können.

Im Rahmen unserer allgemeineren Theorie einer netzwerkbasierten Wissenschaftsgeschichte, die wir am Anfang geschildert haben, sind diese Forschungsprogramme ein erster Ansatzpunkt für die Konstruktion eines semantischen Netzwerkes. Wir können hier nur die ersten Ansätze des Umganges mit diesem Netzwerk beschreiben. Bisher haben wir nur die Zuordnung von Personen und Forschungsprogrammen näher untersucht und dies auch nur für einen eingeschränkten Zeitraum. Zum Wechselspiel der Forschungsprogramme untereinander, die die wesentliche Struktur des semantischen Raumes bestimmen würden, haben wir noch kein Entwicklungs- bzw. Abhängigkeitsmodell erstellt. Die im Folgenden geschilderten Ansätze sind in diesem Zusammenhang die Dokumentation der Indiziensuche auf dem Weg zu einer modelltheoretischen Beschreibung. Die Beziehungen, die wir an dieser Stelle letztendlich suchen, entsprechen den Bausteinen eines *epistemischen Handlungsraumes*, den wir beispielhaft in Abschnitt 5.11 beschrieben haben. Zu dieser Indiziensuche gehören auch die in Abschnitt 7.7 noch zu schildernden bibliometrischen Untersuchungen.

Auf der ersten Ebene betrachten wir die Arbeit in einem Forschungsprogramm als Eigenschaften der Personen des Personennetzwerkes. Die Frage, der wir im Folgenden empirisch nachgehen, ist die Rolle der Forschungsarbeiten im Bereich der kosmischen Strahlung (Radiation, RA18). Die Hypothese ist, dass diese ein wesentlicher Beitrag für die Ausprägung des weiteren Forschungsfeldes der ART gewesen sind. Daher schauen wir uns die Rolle dieses Bereiches genauer an. Zunächst betrachten wir den Zusammenhang zwischen Kooperationen und Forschungsfeld. Betrachten wir die absoluten Zahlen und addieren für alle Jahre die Kooperationen auf,[36] dann liegt tatsächlich RA18 im oberen Bereich nur übertroffen von bereits per Definition übergreifenden Forschungsfeldern: Mathematik (RA10), Grundlagen (RA7), Einzelfälle (RA15)[37], Quantentheorie (RA17) (Abb. 7.32).

[36] Hierbei werden Kooperationen zwischen Personen, die über mehrere Jahre laufen, mehrfach gezählt.

[37] Einzelfälle sind hierbei in dem Sinne überlappend, als dass hierzu eine breite Kategorie von Problemfeldern gehört.

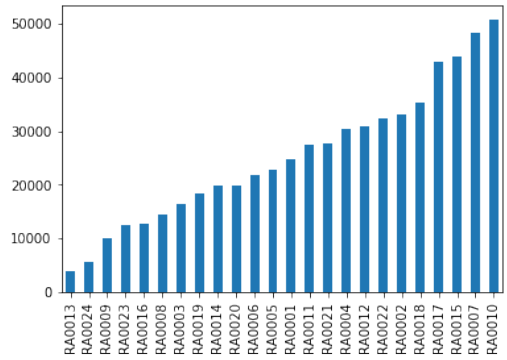


Abbildung 7.32: Anzahl der Kooperationen mit anderen Forschungsfeldern.

Auch in der Heatmap (Abb 7.33) ergibt sich ein ähnliches Bild.

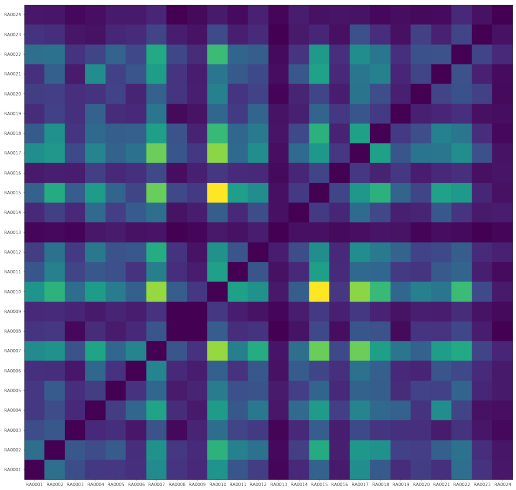


Abbildung 7.33: Heatmap: Kooperationen mit anderen Forschungsfeldern.

ERGM

Bei den Untersuchungen des Netzwerks mit *ERGM*[38] ist es uns bisher nicht gelungen, Annahmen zu formulieren, die zu einem nicht degenerierten Netzwerk führen, wenn wir die entsprechenden Forschungsagenden mit einbeziehen. Wir sehen aber in Abb. 7.1 zumindest eine Tendenz für Nationalitäten und die gröbere disziplinäre Aufteilung.

Bessere Ergebnisse erhalten wir, wenn wir uns lediglich auf die Jahre 1955-1970 beziehen; hier bekommen wir zusätzlich statistisch-signifikante Aussagen für die Forschungsfelder. Vergleichen wir die Jahre 1965-1980 (7.3),1955-1970 (7.4) und 1950-1965 (7.5), so bekommen wir einen ersten Hinweis, dass tatsächlich in der Frühphase sowie in der späteren Phase die Arbeit an kosmischer Strahlung zu einer Verdichtung des Netzes geführt hat. Wir haben für die frühere Periode und spätere Phase eine positive Korrelation, während wir in der zweiten Phase sogar eine negative Korrelation erhalten.[39] Als Modell benutzen wir für diese beiden Berechnungen:

[38] Siehe Abschnitt 5.6.

[39] Wie bereits angemerkt gilt dieses unter der erheblichen Einschränkung, dass das Modell nicht konvergiert.

	Dependent variable:	gr_nw
edges	−6.102***	(0.150)
nodematch.RA0018	−0.032	(0.052)
nodematch.disc.UNKNOWN	−Inf.000***	(0.000)
nodematch.disc.astronomer	−Inf.000***	(0.000)
nodematch.disc.astrophysicist	1.078***	(0.168)
nodematch.disc.cosmologist	−Inf.000***	(0.000)
nodematch.disc.mathematician	0.959***	(0.069)
nodematch.disc.philosopher	−Inf.000***	(0.000)
nodematch.disc.physicist	0.439***	(0.039)
nodematch.nat.	1.852***	(0.521)
nodematch.nat.British	−Inf.000***	(0.000)
nodematch.nat.american	−Inf.000***	(0.000)
nodematch.nat.austrian	1.536**	(0.779)
nodematch.nat.belgian	1.674	(1.243)
nodematch.nat.brazilian	−Inf.000***	(0.000)
nodematch.nat.british	1.303***	(0.138)
nodematch.nat.bulgarian	−Inf.000***	(0.000)
nodematch.nat.canadian	−Inf.000***	(0.000)
nodematch.nat.czech	2.798	(1.765)
nodematch.nat.danish	−Inf.000***	(0.000)
nodematch.nat.dutch	−Inf.000***	(0.000)
nodematch.nat.filipino	−Inf.000***	(0.000)
nodematch.nat.finnish	−Inf.000***	(0.000)
nodematch.nat.french	1.284***	(0.106)
nodematch.nat.german	1.324***	(0.133)
nodematch.nat.german (DDR)	−Inf.000***	(0.000)
nodematch.nat.german (east)	−Inf.000***	(0.000)
nodematch.nat.german (west)	−Inf.000***	(0.000)
nodematch.nat.german-american	−Inf.000***	(0.000)
nodematch.nat.greek	−Inf.000***	(0.000)
nodematch.nat.hungarian	−Inf.000***	(0.000)
nodematch.nat.indian	1.553***	(0.354)
nodematch.nat.indian, probably us-american?	−Inf.000***	(0.000)
nodematch.nat.irish	3.264	(9.104)
nodematch.nat.israeli	2.540**	(1.269)
nodematch.nat.italian	1.697***	(0.362)
nodematch.nat.italian or argentinian (home country)	−Inf.000***	(0.000)
nodematch.nat.japanese	2.201***	(0.379)
nodematch.nat.mexican	−Inf.000***	(0.000)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Tabelle 7.1: Graph der Jahre 1950-1980 ohne institutionelle Beziehungen. Teil I (Fortsetzung 7.2)

	Dependent variable:	gr_nw
nodematch.nat.new-zealand	−Inf.000***	(0.000)
nodematch.nat.polish	2.700***	(0.683)
nodematch.nat.russian	2.354***	(0.400)
nodematch.nat.south african	−Inf.000***	(0.000)
nodematch.nat.spanish	5.568	
nodematch.nat.swedish	5.351	
nodematch.nat.swiss	5.614	(3.972)
nodematch.nat.syrian	−Inf.000***	(0.000)
nodematch.nat.us-american	0.956***	(0.050)
nodematch.nat.yugoslavian	−Inf.000***	(0.000)
nodematch.RA0017	−0.053	(0.047)
nodematch.RA0010	0.115***	(0.042)
nodematch.RA0013	0.109	(0.095)
nodematch.RA0024	0.051	(0.083)
nodematch.RA0009	−0.104*	(0.053)
nodematch.RA0007	−0.305***	(0.048)
nodematch.RA0015	0.032	(0.050)
gwesp.fixed.0.2	2.077***	(0.072)
Note:	*p<0.1; **p<0.05; ***p<0.01	

Tabelle 7.2: Graph der Jahre 1950-1980 ohne institutionelle Beziehungen. Teil II (Fortsetzung von 7.1)

```
md = ergm(gr_nw ~ edges + nodematch("RA0018") +
nodematch("disc") + nodematch("nat") +
nodematch("RA0017") + nodematch("RA0010") +
nodematch("RA0013") + nodematch("RA0024") +
nodematch("RA0009") +
nodematch("RA0007") + nodematch("RA0015") +
         gwesp(0.2,fixed=TRUE), control = control)
```

	Dependent variable:	gr_nw
edges	−8.280***	(0.260)
nodematch.RA0018	0.181***	(0.040)
nodematch.disc	0.161***	(0.042)
nodematch.nat	0.581***	(0.047)
nodematch.RA0017	0.124***	(0.042)
nodematch.RA0010	0.257***	(0.046)
nodematch.RA0013	0.015	(0.060)
nodematch.RA0024	0.195**	(0.098)
nodematch.RA0009	−0.041	(0.049)
nodematch.RA0007	−0.116***	(0.042)
nodematch.RA0015	0.108***	(0.042)
gwesp.fixed.0.2	3.952***	(0.182)
Akaike Inf. Crit.	3,072.584	
Bayesian Inf. Crit.	3,170.612	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Tabelle 7.3: Graph der Jahre 1965-1980 ohne institutionelle Beziehungen.

SIENA

Nehmen wir die Graphen 1950 -1955,1955 -1960,1960 -1965,1965 -1970 als Wellen für SIENA.[40] Für das Modell nehmen wir an, dass die Arbeit mit kosmischer Strahlung eine engere Kooperation auslöst. Im Sinne der Modellbildung von RSIENA heißt dies, dass wir den Effekt erwarten, dass Personen, die in dem Feld aktiv sind, in das Netzwerk einbezogen werden, d.h. wir erwarten eine *covariate related popularity* [222, S. 371]. Außerdem vermuten wir, dass diese Personen im Netz aktiver sind (*covariate related activity*). Im Sinne der allgemeineren Idee, dass in den späten 50er Jahren und in den 60er Jahren eine Verdichtung des Netzes im Sinne der Annahmen von Bettencourt und Kaiser[41] stattfindet, vermuten wir außerdem *Popularitäts-* und *Transitivitätseffekte* [222, S. 370]. Und in der Tat geben die Daten eine grobe Bestätigung der Vermutung über die Rolle der Strahlungsforschung. In der Phase des Anstiegs der Aktivitäten im Bereich der ART nach der Low-Water-Mark Periode

[40] Siehe Abschnitt 5.7.
[41] Siehe Abschnitt 5.9.

	Dependent variable:	gr_nw
edges	−6.095***	(0.119)
nodematch.RA0018	0.050	(0.031)
nodematch.disc	0.552***	(0.022)
nodematch.nat	1.172***	(0.019)
nodematch.RA0017	0.106***	(0.028)
nodematch.RA0010	0.153***	(0.035)
nodematch.RA0013	0.204***	(0.049)
nodematch.RA0024	−0.071	(0.052)
nodematch.RA0009	−0.084*	(0.043)
nodematch.RA0007	−0.394***	(0.045)
nodematch.RA0015	0.091***	(0.034)
gwesp.fixed.0.2	2.294***	(0.063)
Akaike Inf. Crit.	3,612.134	
Bayesian Inf. Crit.	3,719.819	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Tabelle 7.4: Graph der Jahre 1955-1970 ohne institutionelle Beziehungen.

	Dependent variable:	gr_nw
edges	−6.372***	(0.251)
nodematch.RA0018	0.108*	(0.057)
nodematch.disc	0.256***	(0.041)
nodematch.nat	0.800***	(0.054)
nodematch.RA0017	−0.003	(0.052)
nodematch.RA0010	0.127***	(0.049)
nodematch.RA0013	−0.072	(0.077)
nodematch.RA0024	−0.014	(0.060)
nodematch.RA0009	−0.022	(0.075)
nodematch.RA0007	−0.319***	(0.061)
nodematch.RA0015	0.064	
	(0.052)	
gwesp.fixed.0.2	2.615***	(0.189)
Akaike Inf. Crit.	2,995.756	
Bayesian Inf. Crit.	3,098.246	
Note:	*p<0.1; **p<0.05; ***p<0.01	

Tabelle 7.5: Graph der Jahre 1950-1965 ohne institutionelle Beziehungen.

sehen wir eine positive Korrelation der Transitivität für das Forschungsfeld (Abb. 7.6). In der Phase danach sehen wir nun eine Änderung der Verhältnisse. In dieser Periode führt Forschung im Bereich der Strahlung zwar zu höherer Popularität, aber nicht mehr zu einer Verdichtung in diesem Bereich (Abb. 7.7), wobei der statistische Fehler des Modells jedoch hoch ist.

Rate parameters:						
0.1		Rate parameter period 1	0.42	(0.06)		
0.2		Rate parameter period 2	0.68	(0.09)		
0.3		Rate parameter period 3	1.06	(0.12)		
0.4		Rate parameter period 4	0.88	(0.09)		
Other parameters:						
1.	eval	degree (density)	-3.56	(0.12)	-0.06	
2.	eval	transitive triads	1.26	(0.48)	0.05	
3.	eval	ra.18_coVar similarity	-0.77	(0.17)	0.06	
4.	eval	ind. pop.^(1/1) weighted ra.18_coVar	-0.28	(0.18)	-0.09	
5.	eval	transitive triads same ra.18_coVar	1.32	(0.35)	0.07	

Tabelle 7.6: RSiena Simulation für die Jahre 1955, 1958, 1961, 1964, 1967

Rate parameters:						
0.1		Rate parameter period 1	1.06	(0.13)		
0.2		Rate parameter period 2	0.96	(0.12)		
0.3		Rate parameter period 3	0.73	(0.09)		
0.4		Rate parameter period 4	0.52	(0.09)		
Other parameters:						
1.	eval	degree (density)	-4.73	(0.22)	-0.02	
2.	eval	transitive triads	3.27	(0.67)	0.03	
3.	eval	ra.18_coVar similarity	-0.18	(0.21)	0.01	
4.	eval	ind. pop.^(1/1) weighted ra.18_coVar	0.38	(0.19)	-0.05	
5.	eval	transitive triads same ra.18_coVar	-0.18	(0.41)	0.02	

Tabelle 7.7: RSiena Simulation für die Jahre 1965, 1968, 1971, 1974, 1977

7.6 Zeitliche Entwicklung des Graphen

Um einen besseren Eindruck von der zeitlichen Entwicklung zu bekommen, ermitteln wir weitere Parameter, die eine Einschätzung der Wachstumsdynamik des Netzes zulassen. Hier steht die Frage nach der Dynamik im Vergleich mit der Entwicklung von Graphen unter Annahme unterschiedlicher Modelle des Wachstums, sowie im Vergleich zu anderen Annahmen zur Etablierung von wissenschaftlichen

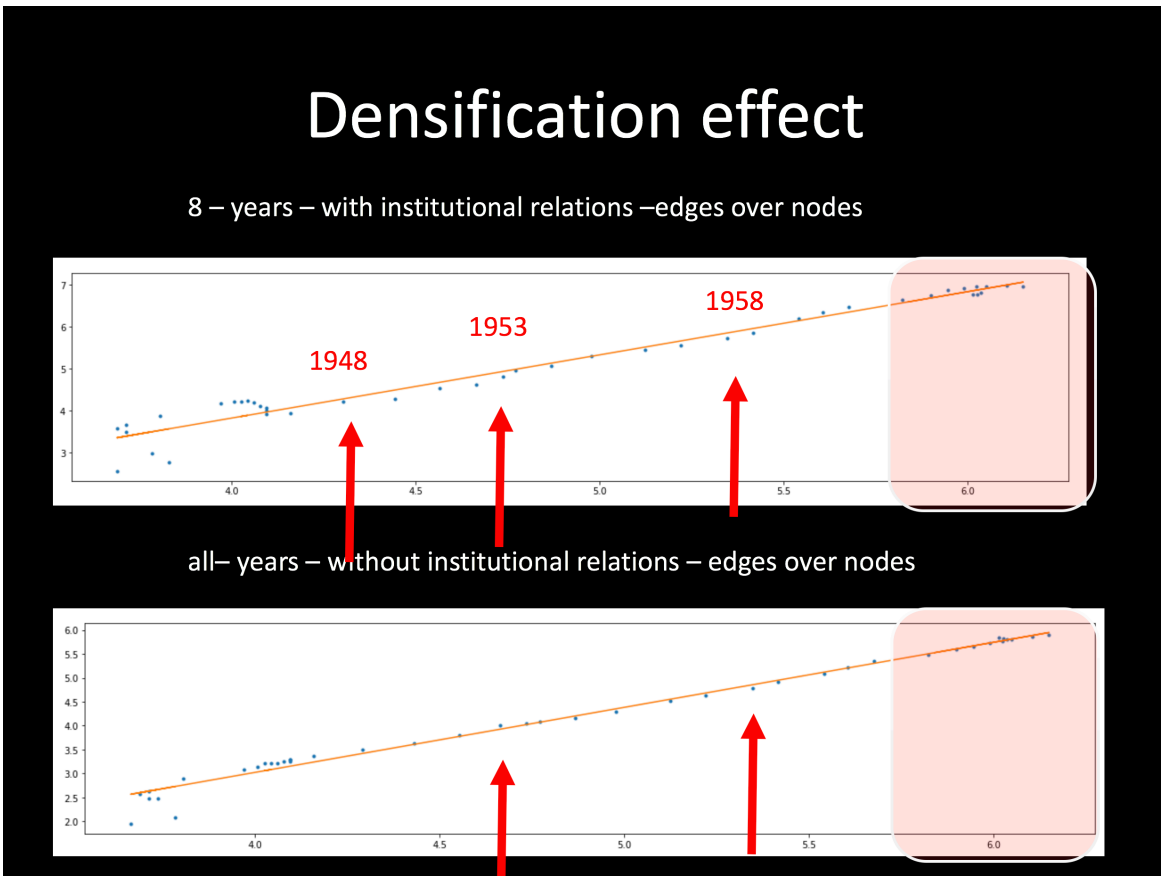


Abbildung 7.34: Densifikationseffekt für zwei Beispiele

Disziplinen im Vordergrund. Die folgende Zusammenstellung dient hier zur Einführung in die Methoden. Eine ausführliche Diskussion und wissenschaftshistorische Einschätzung der Ergebnisse würde den Rahmen dieser Arbeit sprengen.

7.6.1 Etablierung wissenschaftlicher Felder

Bettencourt und Kaiser diskutieren in einer Reihe von Artikeln [23, 22, 21] die Veränderung der Struktur von Kooperationsnetzwerken. Kooperation wird hier bestimmt durch eine Analyse der Kozitationen. Ihre Annahme ist hierbei, dass das Verhältnis von Knoten und Kanten sich skalenfrei mit einer Steigung größer als 1 entwickelt, wenn sich ein Feld etabliert hat, sowie dass sich der Radius der größten Komponenten auf einem Niveau einpendelt. In unserer Studie betrachten wir die Entwicklung des Netzwerkes der Kooperationsbeziehungen für unterschiedliche Parameter des Nachhallfaktors, sowohl einmal unter Berücksichtigung der institutionellen Bindungen als auch ohne diese. Abbildung 7.34 zeigt diesen Verdichtungseffekt für zwei Fälle. In beiden Fällen bemerken wir strukturelle Effekte und eine Annäherung der Kurven an eine Gerade von etwa dem Jahre 1948 an, danach ergeben sich noch einmal leichte Veränderungen der Steigung in den Jahren 1953 und 1958. Nach 1970 sind die Daten nicht mehr eindeutig zu interpretieren.[42] Die Entwicklung des Radius der größten Komponente ist nicht eindeutig zu bewerten. Die bisherigen Daten zeigen das Erreichen eines Plateaus in zwei

[42] *Dynamische Entwicklung der Graphen - Bettencourt.pynb*

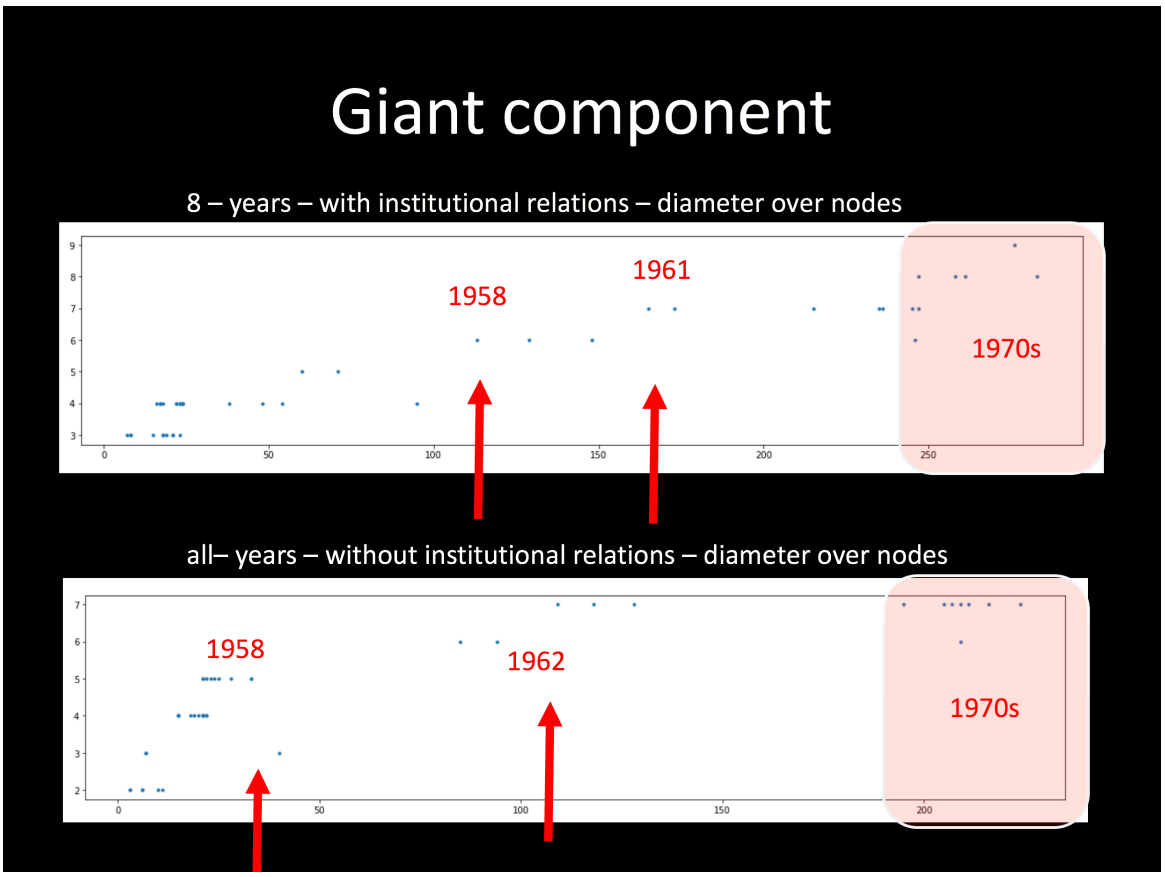


Abbildung 7.35: Entwicklung des Radius der größten Komponente. Mit wenigen Abweichungen zeigt die Kurve das von Bettencourt und Kaiser vorhergesagte Muster, wenn institutionelle Beziehungen nicht berücksichtigt werden. Im anderen Falle gibt es mehrere deutliche Sprünge.

bzw. drei Wellen mit Sprüngen in den Jahre 1958 und 1961/1962. Auch hier gilt wieder, dass die 1970er Jahre sich zumindest bei Berücksichtigung institutioneller Beziehung nicht mehr einordnen lassen.

7.6.2 Einordnung der Entwicklung im Rahmen der Theorie sich entwickelnder Graphen

Dynamisch wachsende Graphen werden systematisch seit mehreren Jahren untersucht. In [120, Teil II und Teil IV] findet sich ein guter Überblick über die unterschiedlichen Annahmen und die sich daraus ergebenen Charakteristiken für das Wachstum von Netzwerken. Wir untersuchen in *Degree_distributions.ipynb* nur einige wenige Merkmale. So betrachten wir die *Degree-Verteilung* für den Gesamtgraphen unter Annahme des *Hybridmodells*[43] sowie die Entwicklung des mit diesem Modell verbundenen Parameters α, der den Anteil von Verbindungen im Sinne einer bevorzugten Auswahl (*preferential attachment*) gegenüber eine zufälligen Auswahl bestimmt. Hierbei bedeutet $\alpha = 1$ vollständig zufällige Wahl, während $\alpha = 0$ vollständig bevorzugte Auswahl bedeutet. Für den vollständigen Graphen erhalten wir hier $\alpha = 0.5463 \pm 0.0011$ und damit eine Verteilung, bei der sich beide

[43] Siehe Abschnitt 5.5.3.

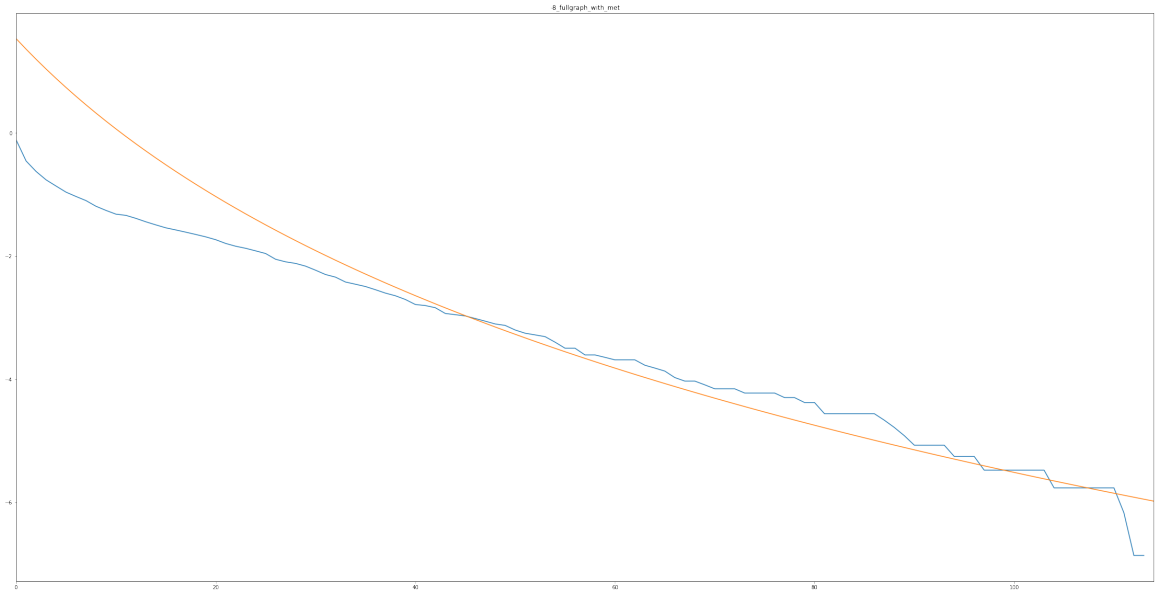


Abbildung 7.36: Dichteverteilung unter Annahme eines Hybridansatzes.

Ansätze die Waage halten. Abbildung 7.36 zeigt den entsprechenden Verlauf der Dichteverteilung mit den realen Daten (blau) und dem Fit nach der mit Gl. 5.2 beschriebenen Hauptfeldnäherung.

7.7 Kozitationsanalysen

Chaomai Chen stellt in [41] einen Ansatz vor, der auf der Grundlage von Kozitationsanalysen Einblicke in Wissenschaftsdynamik erlaubt. Wir haben diesen in Abschnitt 5.10 bereits vorgestellt. Die Daten für die Analyse in dieser Studie stammen aus *Web Of Science* und einer systematischen Schlagwortsuche nach einer Reihe von Begriffen aus dem Bereich der ART. Abbildung 7.37 zeigt die Ergebnisse der Auswertung. Die Bezeichnungen für die Cluster ergeben sich mittels eines Burstness-Algorithmus direkt aus den für den entsprechenden Cluster relevantesten Artikel. Sie geben weitestgehend im Kontext der Entwicklung der ART erwarteten Themenfelder wieder. In der zeitlichen Auflösung ergibt sich hier ebenfalls die vorhergesagte Reihung.

7.8 Zusammenfassung

Die Ergebnisse der Studien zeigen einen Verlauf der Entwicklungen, der in Übereinstimmung mit unseren Vorannahmen ist. Einzelne Effekte, wie das vorzeitige Abfallen der Intensität von Kooperationen vor dem Zweiten Weltkrieg, zeigen sich deutlicher als erwartet. Kozitationsanalysen ergeben zugleich eine überraschend gute Übereinstimmung der Entwicklung von Themenfeldern mit den bisherigen Erklärungen und damit eine erneute Bestätigung des von Chen verfolgten Ansatzes bei der Kozitationsanalyse. Wir sehen Tendenzen, die auf eine beginnende stärkere Rolle von Institutionen ab dem Ende der 1960er Jahre hinweisen. Insbesondere zeigt diese Fallstudie jedoch, dass mit relativ geringen Anpassungen bestehender Verfahren und Hilfsmitten quantitative Aussagen, die

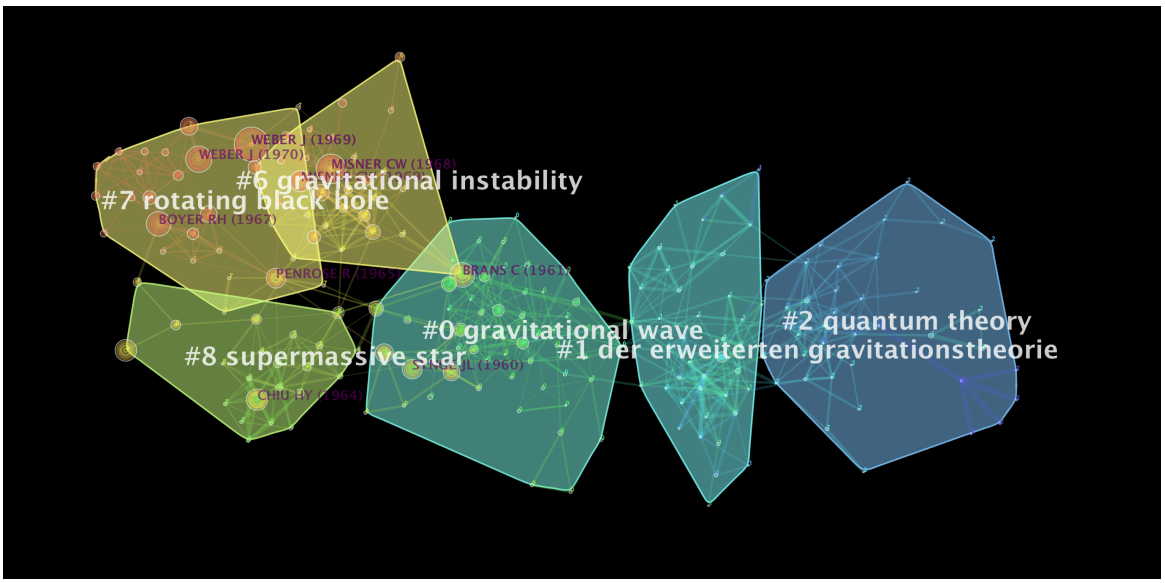


Abbildung 7.37: Mit Citespace erstellte thematische Cluster für Artikel zur ART

Ausgangspunkte für weitergehende historische Studien sein können, gewonnen werden können. Für die nächste Stufe des Projektes sehen wir daher eine detaillierte Studie der Themenentwicklung vor und planen insbesondere die vorhandenen Volltexte des *Astrophysics Data System* [3] auszuwerten. Dafür existieren Pythontools [4, 39], die ein Textmining erleichtern. Auf dieser Grundlage gehen wir davon aus, mit den beschriebenen Methoden weitere Erkenntnisse über die Entwicklung von Themenfeldern erzielen zu können. Insbesondere die noch relativ oberflächlichen Untersuchungen in 7.5 und die nur auf *Web Of Science* beruhenden Untersuchungen in 7.7 sollen damit auf eine breitere Grundlage gestellt werden. Wir sind uns auch bewusst, dass dafür eine wesentlich detailliertere Modellierung der Dynamiken notwendig ist, um auf der Grundlage von statistischen Modellen Aussagen zu treffen. Insbesondere benötigen wir noch verlässliche Modelle, die Aussagen über die Entstehung und die Rolle von Institutionen ermöglichen. Jedoch sehen wir in den ersten Ergebnissen eine Bestätigung, dass ein solcher noch wesentlich zeitaufwändigerer Ansatz vielversprechend ist und weiter verfolgt werden sollte.

Vor dem Hintergrund des theoretischen Konzeptes der *epistemischen Netzwerke*[44] haben wir mit dieser Studie eine Komponente des sozialen Netzwerkes – Kooperationen und Institutionen –, sowie eine des *semiotischen Netzwerkes* – Zitations- und Kozitationsanalyse – untersucht. Mittels der Untersuchungen der Forschungsfelder ist außerdem der Weg in Richtung der Beschreibung des *semantischen Netzwerkes* beschritten. Jedoch haben wir bisher lediglich unterschiedliche Teilkomponenten unseres komplexen Netzwerkes untersucht und stehen nun vor der Herausforderung, die unterschiedlichen Fragmente zu einem schlüssigen Bild zusammenzuführen.

[44] Siehe Abschnitt 5.13.

Kapitel 8

Strukturen und Netzwerke – Arbeit mit den Daten eines großen historischen Forschungsprojektes

Diese Fallstudie widmet sich einem Teilprojekt innerhalb eines groß angelegten Forschungsprojektes zur Geschichte der Max-Planck-Gesellschaft (MPG). In diesem Projekt arbeitet ein internationales Team von Historikern, Sozialwissenschaftlern und Naturwissenschaftlern an einer übergreifenden Geschichte der MPG. Ziel ist hierbei nicht, eine Sammlung von Einzelgeschichten der Institute der MPG zu erstellen, sondern strukturelle Voraussetzungen für Innovationen und die Auswahl von Forschungsfeldern zu untersuchen. Dieser Ansatz macht es notwendig, organisations- und sozialhistorische Fragestellungen eng mit der Untersuchung der wissenschaftlichen Entwicklung zu verbinden. Die Auswertungen dieses Kapitels beziehen sich auf den Datenbestand von Mitte 2017.

Eine der Fragestellungen ist die nach den treibenden Kräften, die die Prozesse in einer komplexen Organisation bestimmen. Welche Randbedingungen beeinflussen die Formation von Akteursgruppen, welche Themenfelder werden wann aufgegriffen, und wie formieren sich disziplinäre, multi- und transdisziplinäre Arbeitsstrukturen innerhalb der MPG? Weitere Fragen sind, ob Zentralisierungstendenzen, die generell angenommen werden, in den Netzwerkdaten ersichtlich sind und ob sich Tendenzen ablesen lassen, die eine Öffnung der Strukturen im Hinblick auf die Einbeziehung von Akteuren erkennen lässt. Die im Folgenden geschilderte Fallstudie konzentriert sich hierbei stärker auf die Binnenstrukturen. Wie wir im Folgenden sehen werden, sind diese Vorstudien notwendig, um die Quellenlage zu verstehen und Desiderate in den Quellen einzugrenzen, um dann in einem bevorstehenden zweiten Schritt genauer auf komplexere Fragen einzugehen.

Als Arbeitshypothese und als Instrument zur Strukturierung der Herangehensweise werden die mehr als 100 zu untersuchenden Institute in themenspezifische Cluster[1] aufgeteilt. Innerhalb der MPG selbst existiert offiziell nur eine Aufteilung in drei Sektionen: die Biologisch-Medizinische Sektion (BMS), die Chemisch-Physikalisch-Technische Sektion (CPTS) und die Geistes-, Sozial- und Humanwissenschaftliche Sektion (GSHS). Eine zweite Hypothese ist die Annahme von vier unterschiedlichen

[1] Siehe Abschnitt 8.6.2.

historischen Phasen: der Gründungsphase, einer Aufbauphase, eine Phase der Stagnation in Bezug auf die Expansion der MPG und schließlich einer neuen dynamischen Wachstumsphase im Zuge der deutschen Wiedervereinigung.

Die folgende Fallstudie beschäftigt sich hierbei schwerpunktmäßig mit der Frage, inwieweit sich Cluster und Periodisierung in den Gremien der MPG wiederfinden. Die Organisationsstruktur der MPG besteht aus einer Hauptverwaltung, einem Präsidenten, mehreren Vizepräsidenten oder -präsidentinnen sowie den Vorsitzenden der Sektionen auf der einen Seite und andererseits den Instituten, die eine weitreichende Autonomie besitzen. Entscheidungen werden von verschiedenen Gremien gefällt, wobei ein entscheidendes Element hier die Selbstverwaltung der MPG ist: wesentliche Entscheidungen fällt der Wissenschaftliche Rat, der aus allen wissenschaftlichen Mitgliedern der MPG besteht, auf der Hauptversammlung bzw. die Versammlung der Mitglieder der Sektionen, sofern es nur diese betrifft. Die Institute wiederum können eine unterschiedliche Anzahl von Direktoren haben, die in der Regel jeweils eine Abteilung leiten. Das Institut als Ganzes wird kollegial geführt. Hinzu kommen eine Reihe von auswärtigen wissenschaftlichen Mitgliedern, die in der Regel Professuren an Universitäten innehaben und zwar keine Abteilungen leiten, jedoch am Diskussions- und Entscheidungsprozess der Institute und der gesamten MPG beteiligt sind. Hinzu kommt der Senat der MPG, der die formale Entscheidungsgewalt besitzt und mit Vertretern der Länder, des Bundes und der MPG besetzt ist. In der Regel entscheidet dieser nicht gegen den Willen der Sektionen. Der Senat setzt sich aus Mitgliedern der MPG sowie Vertretern aus Politik und Gesellschaft zusammen. An der Spitze der Verwaltung steht die Generalsekretärin oder der Generalsekretär. Wirtschaftliche Entscheidungen werden im Verwaltungsrat der Gesellschaft getroffen. Die Gesellschaft selbst ist in Form eines gemeinnützigen eingetragenen Vereins organisiert. Diese komplexe Struktur bewirkt eine große Anzahl von Kommissionen zumeist aus wissenschaftlichen Mitgliedern, die die wesentlichen Entscheidungen wie die Neuberufungen von Direktoren, Neugründungen oder Schließungen von Instituten vorbereiten. Die Vermutung liegt nahe, dass sich daher aus der Dynamik dieser Kommissionen Rückschlüsse auf Einfluss und Macht auf der einen Seite, aber auch auf die Formierung von Themenfeldern und deren Kontext zur gesellschaftlichen und wissenschaftlichen Öffentlichkeit ziehen lassen. Wir untersuchen daher mit dem Schwerpunkt auf einem der Cluster, *Astronomie und Astrophysik*, die Rolle der großen Zahl dieser unterschiedlichen Kommissionen für die Formation von Interessengruppen und Forschungsfeldern. Dieses Vorhaben dient im Kontext des Projektes zur Geschichte der Max-Planck-Gesellschaft als ein exploratives Beispiel für die Anwendung digitaler Methoden zur Analyse der Dynamik dieses Feldes. Auch hier gilt, dass diese Arbeiten nur im Rahmen eines Teams von Wissenschaftlerinnen und Wissenschaftlern möglich sind.

8.1 Quellen

Wie in jedem historischen Projekt hängen Erfolg und Misserfolg nicht zuletzt von der Tragfähigkeit der Quellen ab. In diesem Fall stellen die Quellengrundlage und der zeithistorische Bezug das Projekt vor Herausforderungen, die in der historischen Forschung und insbesondere in der wissenschaftshistorischen Betrachtung bisher nicht systematisch angegangen worden sind. Als wissenschaftshistori-

sches Projekt steht es in der historisch-kritischen Tradition. Die genaue Analyse der vorliegenden Quellen ist eine der zentralen Herausforderungen. Der relativ lange Zeitraum, die Größe der zu betrachtenden Organisation und ihre umfangreichen internationalen und nationalen Verflechtungen mit anderen Wissenschaftsorganisationen und Forschungseinrichtungen sowie Akteuren aus Politik und Gesellschaft und insbesondere die multidisziplinäre Struktur der MPG erzwingen die Sichtung und Analyse eines großen heterogenen Quellenkorpus, der von administrativen Dokumenten, publizierten und unpublizierten Forschungsergebnissen bis hin zu Zeitzeugeninterviews reicht. Die Daten selbst sind hierbei eine Mischung aus hochstrukturierten Daten, die sich aus der individuellen Analyse von Dokumenten ergeben und in einer komplexen Datenbank dokumentiert werden, sowie den im Rahmen eines umfangreichen Digitalisierungsprojektes erstellten digitalen Faksimiles von Dokumenten, die erst in maschinell auswertbare Form überführt werden müssen. Letztendlich werden mehrere Kilometer Aktenbestände im Laufe des Projektes digitalisiert werden. Ohne die Unterstützung durch digitale Auswertungsmethoden ist eine Sichtung dieser Quellen nicht realisierbar.

Die Bestände sind hierbei nur in sehr geringem Maße archivalisch bearbeitet und exakt katalogisiert, zumeist handelt es sich um Aktenbestände der Registratur und Bestände aus dem Archiv und den Instituten der MPG, die nur sehr grob gesichtet wurden. Hinzu kommen Publikationsdaten, wissenschaftliche Artikel und digitale Datenbestände aus den Fachdisziplinen. Erschließung, Digitalisierung und Analyse gehen hierbei Hand in Hand. Zugleich ergeben sich durch die Arbeit an der Erschließung neue Fragestellungen und damit immer wieder die Notwendigkeiten, die Struktur der Datenbank zu erweitern und abzuändern. Aus pragmatischen Gründen werden unterschiedliche Systeme für die Dateneingabe und für die Auswertung benutzt. Metadaten aus unterschiedlichen Quellen müssen daher integriert werden. Die Anforderung, strukturell und inhaltlich höchst unterschiedliche Daten in einem einheitlichen System zur Verfügung zu stellen und insbesondere unter verschiedenen Aspekten zu verknüpfen, macht das Projekt zu einem idealen Kandidaten für einen modellbasierten Ansatz und einen Testkandidaten für Fragen der Datenintegration im Kontext von *Linked Open Data*. Können Datenmodelle gefunden werden, die flexibel genug sind, um neue Datenbestände ohne größeren Aufwand zu integrieren, aber immer noch spezifisch genug sind, um konkrete Forschungsfragen zu beantworten, ohne bei jeder hinzukommenden Datenquelle und neuen Forschungsfrage grundlegend in die Struktur der Datenmodellierung eingreifen zu müssen? Kann das Modell so zugeschnitten werden, dass es Anfragen erlaubt, deren Resultate eine verlässliche historische Interpretation ermöglichen?

8.2 Personen und Kommissionen

Ausgangspunkt für die Analysen der Kommissionen ist die Datenbank, oder genauer: sind die Datenbanken des Projektes zur Geschichte der MPG. Die Ausgangsdatenbanken sind komplexe relationale Datenbanken mit PostgreSQL als Backend. Das Datenmodell ist bestimmt durch Objektmodelle, die als Pythonobjekte in Django[2]beschrieben werden. Dieses Datenmodell ist während der Arbeit mit den Daten gewachsen und zunehmend komplexer geworden. Ausgangspunkt war zunächst die Not-

[2]Siehe 6.4.

wendigkeit, alle innerhalb des Projektes relevanten Akteure zu bestimmen und ihre Zugehörigkeit zur MPG, ihre Rolle innerhalb der MPG sowie die Zuordnung zu den einzelnen Instituten der MPG festzuhalten. Die Quellenlage ist unübersichtlich, und digital bereits vorhandene Datenbestände, die bis in die Anfangstage der MPG zurückreichen, sind nicht vorhanden. Es erwies sich als notwendig, verschiedene nur gedruckt vorhandene Quellen und einzelne bereits vorhandene Personendatenbanken bei der Erfassung zu berücksichtigen und zusammenzuführen. Bereits bei der ersten Grobsichtung zeigten sich Inkonsistenzen und Ungenauigkeiten bei fast allen relevanten Daten, beginnend bei Geburtsdaten, Namensschreibungen bis hin zur Rolle in der MPG. Die Ursachen sind vielfältig und basieren auf tatsächlichen Änderungen von Namensschreibungen oder Wechseln des Nachnamens, etwa aufgrund von Heirat oder Emigration, sowie unterschiedlichen Definitionen von Daten und Datenstrukturen. Die Datenbank soll allen am Projekt Beteiligten weitestgehend zugänglich sein, woraus sich unmittelbar Probleme mit dem Datenschutz ergeben, auf die hier nicht näher eingegangen werden soll, die aber im Projekt und auch in dieser Fallstudie berücksichtigt werden mussten. Die Eingabe selbst erfolgte durch studentische Hilfskräfte und durch Mitarbeiter des Projektes mit unterschiedlichem Hintergrundwissen zu sehr verschiedenen Zeitpunkten. Auch dies führte zu zusätzlichen Fehlerquellen. Aus dieser Problemstellung ergaben sich unmittelbar konkrete Anforderungen an das Datenbankdesign. Für (fast) alle Felder musste es möglich sein, alternative Lesarten und Informationen mit aufzunehmen. Es sollten Informationen als gesichert gekennzeichnet werden können, wobei die Geschichte des Eintrages und vor allem seine Quellengrundlage weitestgehend nachvollziehbar sein sollte. Einerseits aus pragmatischen, andererseits aus inhaltlichen Gründen sind – wie häufig in Projekten innerhalb der digital gestützten historischen Forschung – Datenmodellierung, Entwicklung und Anpassung der Arbeitsumgebung, Datenerfassung und Forschung eng miteinander verflochten, so dass traditionelle Phasenmodelle der Softwareentwicklung nicht unmittelbar anwendbar sind. Im Laufe der Datenerfassung und den sich daraus ergebenen ersten Einsichten in die Daten und daraus neu aufgeworfenen zusätzlichen Fragestellungen ergeben sich zwangsläufig neue Anforderungen an das Datenmodell. Felder, für die zunächst ein Freitextfeld ausreichend erschien, müssen nun stärker kontrolliert werden. Beziehungen zwischen Personen, die festgehalten werden sollten, differenzieren sich weiter aus. Es werden neue Datenquellen erschlossen, die möglichst nahtlos in das Projekt integriert werden sollen. So verschob sich die Funktion der Datenbank, die ursprünglich als rein personenzentrierte Datenbank angedacht war, die lediglich der Standarisierung und dem Überblick dienen sollte, weiter in Richtung eines umfassenden Informationssystems, das auch die Beziehungen der Personen untereinander wiedergeben sollte. Insbesondere stellte sich die Frage nach der Rolle der einzelnen Gremien und hier besonders die der unterschiedlichen Kommissionen immer drängender, so dass in die Datenbank stetig weitere Querreferenzen aufgenommen werden mussten.[3]

Alle diese Anforderungen legen eine Modellierung in RDF und einem Triplestore als Backend nahe. Dieses war von allen an der Softwareentwicklung Beteiligten auch so von Anfang an vorgesehen. Jedoch konnten wir keine RDF-basierte Lösung identifizieren, die vor allem die erforderliche Endbenutzerfreundlichkeit anbot. Nach einer Prototypphase mit FileMaker fiel daher die Entscheidung auf

[3] Gleichzeitig gilt es immer, den datenschutzrechtlichen Verpflichtungen Genüge zu tun.

Django,[4] das in der Tat eine hohe Flexibilität ermöglicht – jedoch zu dem Preis, dass die Struktur der Daten leicht unübersichtlich wird und sich Relationen in Kreuztabellen wiederfinden.

Das Problem, eine graphenbasierte Lösung zu finden, wurde umso drängender, als sich aus der vermuteten Relevanz der Kommission zwangsläufig der Wunsch ergab, die dahinter stehenden strukturellen Veränderungen und Einflusszonen näher analysieren zu können. Die vorhandene Datenstruktur lässt hier zwar prinzipiell das Erstellen eines Abhängigkeits- und Einflussnetzwerkes zu. Jedoch gilt nicht, nur aber gerade für die historische Interpretation, dass die Gewichtung der Beziehungen zwischen einzelnen Personen bzw. Knoten allgemein in einem solchen Netzwerk nicht unmittelbar einsichtig ist und von der Einzelfallstudie und von Annahmen über Funktionsmechanismen abhängt. Um den Anforderungen an die Nachvollziehbarkeit der Annahmen und die Rückkopplung der globalen Analysen an die Quellengrundlagen zur Ermöglichung von historischen Tiefenanalysen nachzukommen, ist eine genaue Beschreibung der Kanten des Netzwerkes notwendig. Zugleich sollten die Auswirkung von Hypothesen über die Rolle von Bindungen und Personen auf die Netzwerkstruktur möglichst interaktiv oder doch zumindest in überschaubarer Zeit untersucht werden können.

Durch die Modellierung in RDF, basierend auf einer moderaten Erweiterung von CRM,[5] lassen sich diese Anforderungen in transparenter Weise erfüllen. Konkret werden die Daten aus der ursprünglichen Datenbank exportiert, dann in RDF umgewandelt und daraus per SPARQL-Abfragen und Transformationskripten in ein Graphformat (*GraphML*) umgewandelt.[6]

8.3 Von der Datenbank zum Netzwerk

Der Entscheidung, von einer relationalen Datenbank auf RDF überzugehen, wurde, wie oben beschrieben, getrieben von konkreten Forschungsfragen an die Datenbank. Die Modellierung hat daher zunächst zum Ziel, die Ausgangsdaten angepasst an die Forschungsfrage in eine dafür zugeschnittene Struktur zu überführen. Jedoch muss eine Erweiterung und Umstrukturierung abhängig von den sich verändernden Fragestellungen aufgrund neuer Daten und der durch ihre Analysen gewonnenen Einsichten flexibel möglich sein. Außerdem ist es notwendig, die Datengrundlage soweit zu dokumentieren, dass die Ergebnisse nachvollziehbar sind – einerseits in Bezug auf die benutzten Verfahren und die zum Zeitpunkt der Analyse vorhandenen digitalen Daten, andererseits durch die Möglichkeit der direkten Überprüfung durch den Zugriff auf die Quellen. Die Forderung der Nachvollziehbarkeit hat zur Konsequenz, dass bewusst auf einen unmittelbaren Import der Daten aus der Datenbank in den Triplestore verzichtet und stattdessen der Datenbankdump in JSON ausgewertet wird. Dieser wird als eigenständiger Datensatz archiviert. So kann auf diesen zu einem festen Zeitpunkt gesicherten Datensatz zugegriffen werden, um die Auswertungsergebnisse zu überprüfen.[7] Die Umwandlung

[4] Siehe Abschnitt 6.4.

[5] Siehe Abschnitt 4.4.

[6] Prinzipiell ließen sich die meisten Transformationen, die in Skripten ausgeführt werden, auch direkt in SPARL bzw. mittels in SPARQL durch Blazegraph zur Verfügung gestellten Erweiterungen auch ohne den Transformationsschritt in Python durchführen. Dieses führt jedoch zu nicht immer leicht überschaubaren Konstrukten in SPARQL.

[7] Dies entspricht dem Verfahren, dass wir auch beim Export der Daten aus FileMaker in der Studie zur ART gewählt haben (Abschnitt 7.3.2). Datenschutzgründe erlauben es leider nicht, den Datensatz des GMPG-Projektes vollständig als

in RDF erfolgt durch eine Reihe von Pythonskripten und -notizbüchern. Auch diese liegen versioniert in einem *Git-Repositorium*.[8]

Im Prinzip könnten auch an dieser Stelle die erstellten RDF-Daten getrennt archiviert werden. Wir haben jedoch aus praktischen Gründen darauf verzichtet, da gegebenenfalls die Daten mit den Skripten der entsprechenden Version und den Daten reproduziert werden können. Diese Daten werden dann in einen Triplestore importiert. Auch hier wäre eine Versionierung über *named graphs* oder analoge Mechanismen möglich. Pragmatisch erwies sich jedoch die damit verbundene „Buchführung“ über die jeweils aktuell gültigen Graphen, insbesondere bei „spontanen“ Anfragen an die Datenbank in SPARQL, als nicht zuverlässig handhabbar, so dass es zu falschen Ergebnissen durch die Einbeziehung veralteter Graphen kommt. Aus dem Triplestore werden mittels Pythonskripten, die im wesentlichen Wrapper von SPARQL-Anfragen sind, Netzwerke erstellt, die dann in *GraphML* ausgegeben werden.[9]

8.3.1 Modellierung

Der Ausgangspunkt für die Modellierung war, die Frage beantworten zu können: Wer sitzt mit wem, wann, in welcher Funktion in einer Kommission und wie stehen Kommissionen untereinander in Beziehung? Außerdem sollten im ersten Schritt zunächst nur die als weitestgehend gesichert geltenden Eigenschaften und Beziehungen abgebildet werden. Erst schrittweise – abhängig von den Forschungsfragen – werden dann weitere Eigenschaften aufgenommen, die sich entweder aus der Interpretation und Bewertung der vorhandenen Daten oder aus der Erschließung neuer Quellen ergeben. Anschaulich umgeben wir die Kerndaten schrittweise mit weiteren Schalen, die mit unterschiedlicher Genauigkeit nähere Informationen über die Objekte des Systems enthalten.

Konkret heißt dieses: Personen aus der Datenbank werden direkt in **E21_Person** von CRM abgebildet und im Wesentlichen werden zunächst keine weiteren Beziehungen angebunden. Die Rolle der Kommissionen ist zunächst nicht unmittelbar offensichtlich. Prinzipiell können sie als eine Form eines **E72_Legal_Object** im Sinne von CRM gedeutet werden, zum jetzigen Zeitpunkt scheint es uns jedoch ausreichend, sie als eine Gruppe von Akteuren und selbst als Akteur zu verstehen. Um uns die Flexibilität einer Umordnung zu ermöglichen, wurden die Kommissionen als Unterklasse von **E74_Group** realisiert.[10] Die Zeitdauer der Mitgliedschaft einer Person in einer Kommission variiert und es ist möglich, dass eine Person in unterschiedlichen Rollen zu verschiedenen Zeitpunkten Mitglied einer Kommission ist.[11] Die Kommissionsmitgliedschaft wird daher als eine eigene Klasse **sr:CommissionMembership** eingeführt, die über **sr:is_part_of_commission** mit der Kommission

OpenData zu publizieren. Die Daten liegen jedoch zugriffseingeschränkt in einem Datenrepositorium vor.

[8] Realisiert ist dieses mittels Apache-Jenkins (siehe Abschnitt 6.7) über gesteuerte Workflows, die einerseits einen „Daily-Build“ auf der Grundlage der jeweils aktuellen Datensätze erstellen, andererseits wohldefinierte Versionen erzeugen können.

[9] Siehe Abschnitt 6.2.3.

[10] Auch eine Zuordnung zu **E40_Legal_Body** wäre denkbar, ist aber im Detail nicht immer zu rechtfertigen, so dass wir in der Oberklasse bleiben. Die Problemstellung entspricht der in 7.2, dort in Bezug auf die Zuordnung von **sr:Institution** geschildert.

[11] Informationen über letztere liegen bisher nur in Form von Freitexteinträgen in der Datenbank vor, so dass dieses noch nicht systematisch erfasst ist.

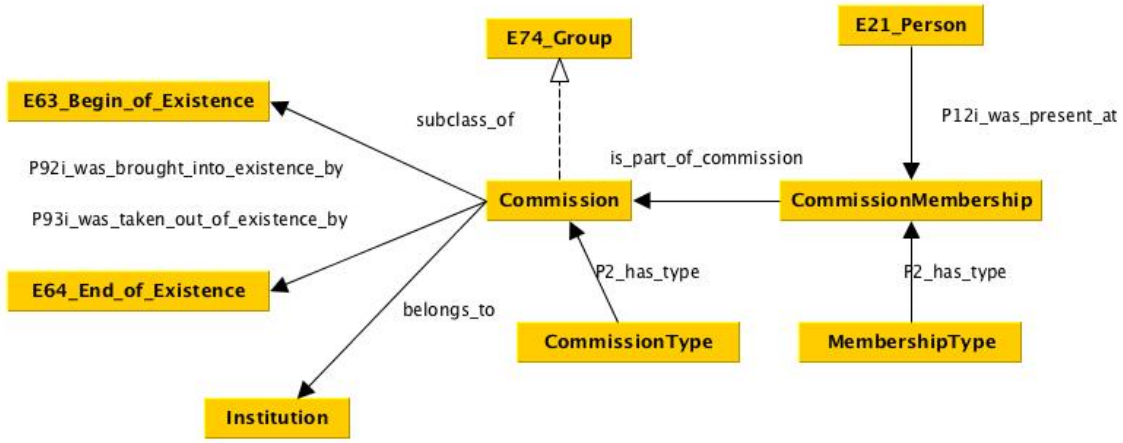


Abbildung 8.1: Kommissionen und Kommissionsmitgliedschaft

verbunden wird. Die Personen selbst sind mit der Kommissionsmitgliedschaft über Relationen vom Type **P12i_was_present_at** verbunden. Die Art der Mitgliedschaft wird in **sr:MembershipType** festgehalten.[12]

Zeitspannen

Für die Kommission liegen häufig explizit Quellen mit genaueren Angaben über die Umstände ihrer Entstehung sowie über Ereignisse, die sich an das Ende einer Kommission anschließen, vor – insbesondere im Falle von Gründungs- und Schließungskommission ist dieses offensichtlich. Für Zukunftskommissionen ist dieses häufig noch eine Frage der Forschung, aber naturgemäß erwarten wir auch hier entsprechende Konsequenzen. Daher wird den Kommissionen keine Zeitspanne direkt zugeordnet, sondern es werden explizit Gründungs- und Schließungsereignisse als Instanzen von **E63_Begin_of_Existence** und **E64_End_of_Existence** beschrieben, diese haben jeweils ein zugeordnetes festes Datum bzw. eine Zeitspanne. Für die Mitgliedschaften reicht es uns zum jetzigen Zeitpunkt aus, Zeitspannen direkt an das Ereignis der Mitgliedschaft anzubinden.[13] Die einzelnen-CBB3 D730 B6F9 4DEF B316 BE6D 4A45 C4E9 FBEF 80C2 Daten liegen in sehr unterschiedlicher Genauigkeit vor. Häufig ist bisher nur die Jahreszahl erfasst. Um Abfragen zu erleichtern, versehen wir die Zeitspannen jeweils mit den genauen Angaben, wenn vorhanden, und zusätzlich mit der Jahreszahl. Hierbei ist die Jahreszahl selbst mittels *rdfs:label* an eine Instanz von **sr:Year** und das Datum in gleicher Form an **sr:Date** angebunden.[14] Wir machen für das Literal selbst zunächst nur die Annahme, dass es sich um eine Zeichenkette handelt, um die Angaben erst einmal eins zu eins, wie in

[12] Dazu später mehr.

[13] Ähnliche Überlegungen hatten wir in Abschnitt 7.3.2 in der Fallstudie zur ART diskutiert.

[14] Auch hier machen wir zunächst keine weiteren Annahmen und übernehmen entweder den ganzen String oder den ersten Teil der Zeichenkette vor einem „-“ Zeichen, davon ausgehend dass bei korrekter Eingabe in die Datenbank entweder im Datum eine Zahl oder eine Zeichenkette der Form YYYY-MM-DD enthalten ist. Im späteren Verlauf betrachten wir dieses genauer und korrigieren mögliche Fehler.

der Datenbank vorhanden, auch in den Triplestore zu übernehmen.[15]

Weitere Eigenschaften der Kommissionen

Zunächst werden nur die Basisdaten aus der Personendatenbank importiert. Dieses sind neben den Daten über die Kommissionen zusätzlich noch Teile der Personendatensätze. Des Weiteren werden die in der Datenbank vorhandenen Verweise auf Quellen aufgenommen, so dass ersichtlich ist, welchen Eintrag sie dokumentieren. Dies sind bisher im Wesentlichen die Gründungen und Schließungen von Kommissionen sowie der Beginn und das Ende von Mitgliedschaften. Dazu benutzen wir über **P70i_is_documented_in** verbundene Instanzen von **E31_Document** bzw. einer Unterklasse **sr:GMPGResource**. Wenn vorhanden, versehen wir diese mit Referenzen auf Online-Quellen.

Abbildung 8.2: Dokumentation

Hierbei dient uns **P128_carries**. Referenzen sind als Unterklassen von **E84_Information_Carrier** definiert. Außerdem erlaubt die Datengrundlage eine Zuordnung der Kommissionen zu den Sektionen der MPG[16] sowie eine Typisierung nach festgelegten Kriterien.[17] Die so in einem ersten Schritt erfassten Daten sind ausreichend, um eine erste Abbildung des Graphen von RDF in eine Netzwerkstruktur vornehmen zu können. Die in den einzelnen Transformationsschritten erzeugten Graphen werden als ein Set *gmpgg:personsCommissionStage1* von Graphen zusammengefasst.[18]

Schätzung von Enddaten und erste Einschränkung

In diesem Stadium haben 9302 von 9388 Einträgen für die Mitgliedschaft in Kommissionen ein Anfangsdatum, jedoch nur 6903 ein Enddatum. Von den Kommissionen sind in den Daten nur 1000 von 1275 Kommissionen mit Anfangsdaten und nur 742 mit Enddaten versehen. Für eine Auswertung der zeitlichen Entwicklung ist dies nicht ausreichend, so dass erste Annahmen für die Kommissionen getroffen werden müssen. Wir können nach Sichtung der Quellen davon ausgehen, dass, auch wenn es keine explizite Nennung der Gründung der Kommission gibt, aus den belegten Mitgliedschaftszeiträumen auf End- und Anfangsdaten geschlossen werden kann. Wir setzen daher das Minimum bzw.

[15] Die alternative Form einer Typisierung erfordert eine Kontrolle und Transformation der Daten und ist Teil eines späteren Schrittes.

[16] Instanzen von **sr:Institution** einer Unterklasse von **E40_Legal_Body is_part_of**.

[17] Instanzen einer Unterklasse von **E55_Type**

[18] Abbildung 12.1 in Kapitel 12 und Abbildung 8.4.

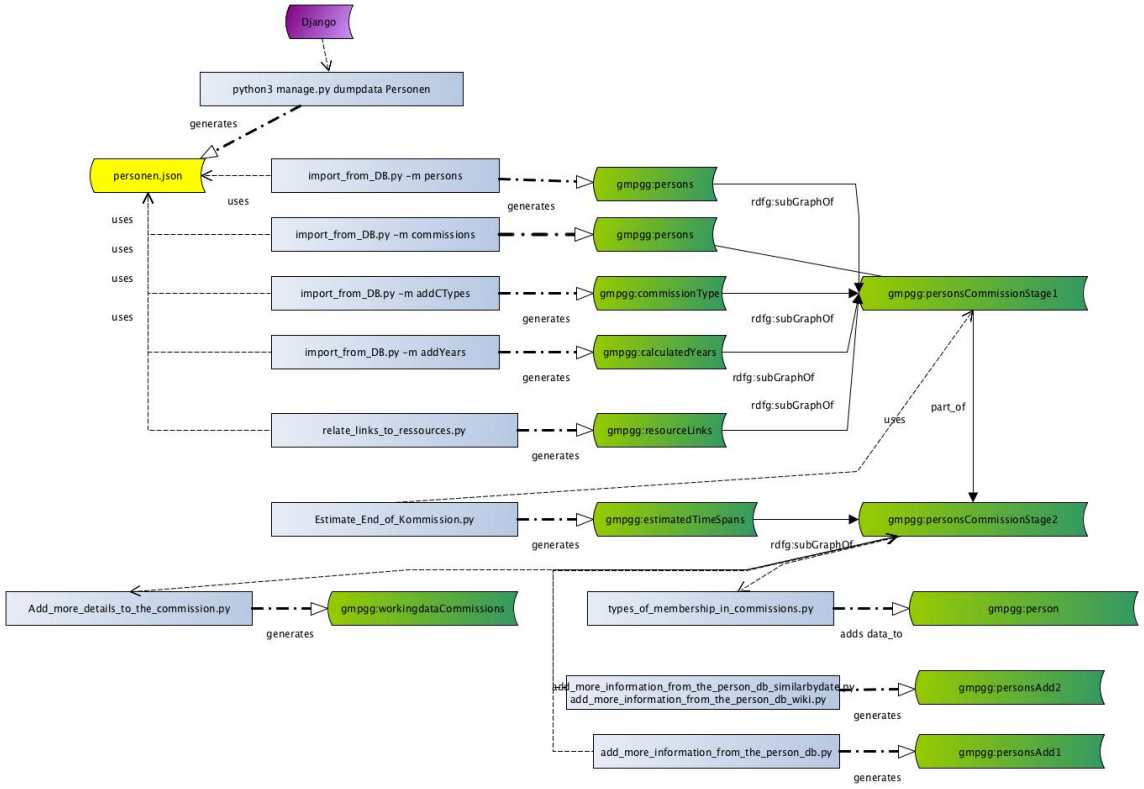


Abbildung 8.3: Übersicht über die Erzeugung der Graphen: die hellblauen Elemente sind hier die benutzten Skripte, grün die Namen der erzeugten Graphen und gelb bezeichnet die Ausgangsdatei.

Maximum der Mitgliedschaftszeiträume als geschätzte Daten für die Kommissionen und versehen die Kommissionen mit einer neuen Zeitspanne als Instanzen von **sr:Estimated_Time-Span**.

Die jetzt immer noch nicht datierbaren Kommissionen sind fast ausschließlich Kommissionen, die nicht in den Kernbereich unserer Fragestellung gehören. Somit gehen wir nun von einer Basis von 1024 Kommissionen aus, davon 1011 datiert.[19] Alle von nun an relevanten *named graphs* sind in einen virtuellen Graphen bzw. einen Graphensatz *gmpgg:personsCommissionStage2* zusammengefasst.[20]

8.3.2 Weitere Reduktionen und neue Attribute

Als Ergebnis der Diskussion um die Bedeutung der Kommissionen steht die Frage nach der unterschiedlichen Relevanz der Typen von Kommissionen im Raum. Als besonders einflussreich werden Berufungs- und Gründungskommissionen sowie Zukunftskommissionen angenommen. Innerhalb der Berufungskommissionen nehmen wir wiederum an, dass die Rolle der Berufung von auswärtigen Mitgliedern an Institute wesentlich geringer ist als die anderer Kommissionen. Diese spezifischen Berufungskommissionen haben zumeist eine sehr kurze Lebensdauer und eher die Form von Ad-hoc-Kommissionen. Im Gegensatz dazu stehen Kommissionen zur Berufung ordentlicher Mitglieder. Deren Zusammensetzung ist das Ergebnis eines längeren Abwägungsprozesses, in den grundsätzliche Überlegungen über die Zukunft des Institutes, an die die neuen Mitglieder berufen werden sollen,

[19] Der Stand für die jeweils aktuellen Daten ist online für die Projektmitarbeiter verfügbar.
[20] Siehe Abbildung 12.2 in Kapitel 12.

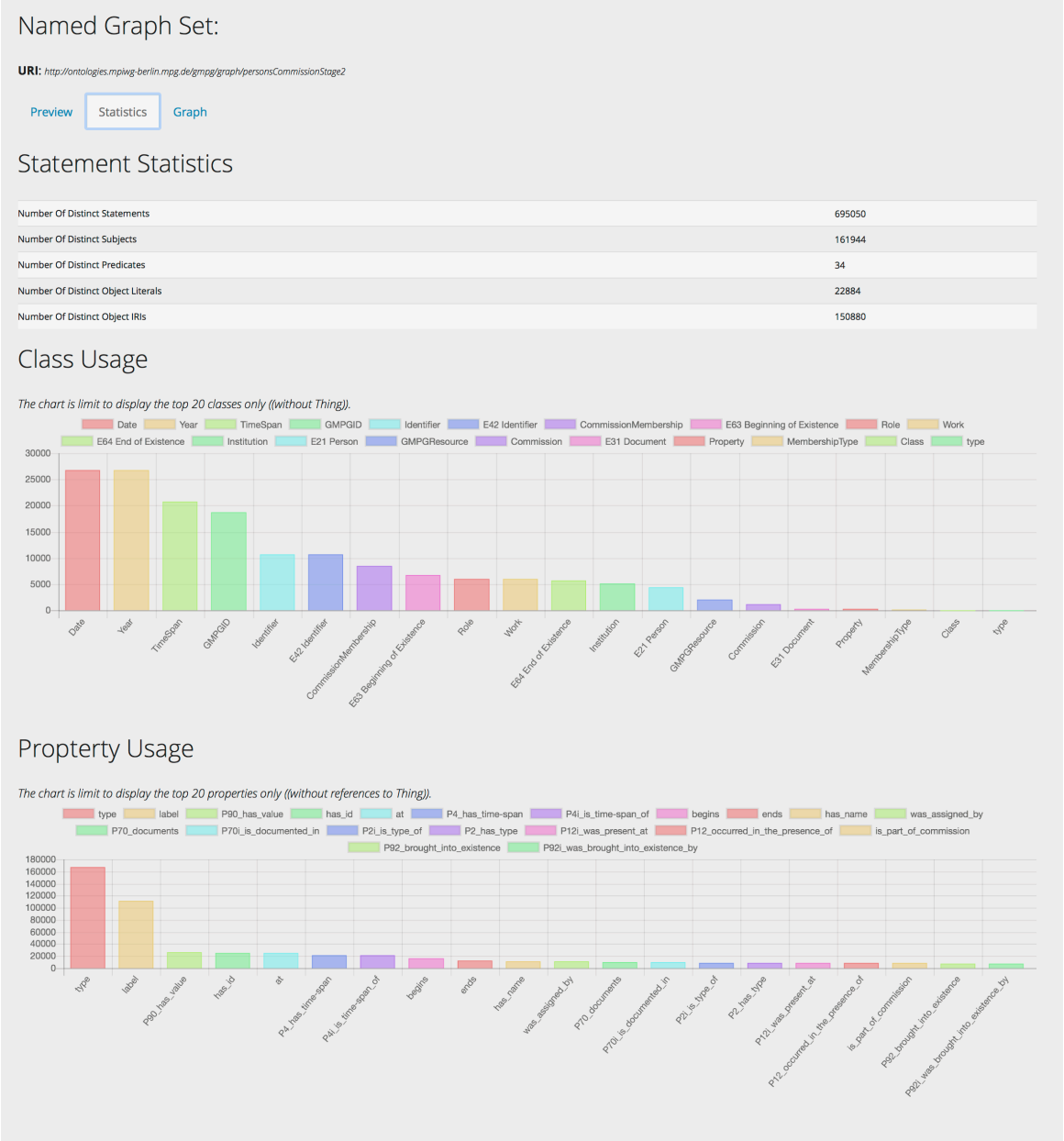


Abbildung 8.4: Übersicht Instanzen und Properties – Personengraph

eingehen. Somit erfordert deren Einsetzung eine inhaltliche Diskussion um die Ausrichtung eines Institutes, während auswärtige Mitglieder in Regel auf Vorschlag eines bereits bestehenden Institutes oder einer Abteilung berufen werden. Die Annahme, dass diese beiden Typen von Kommissionen unterschiedlich bewertet werden sollten, scheint daher angemessen. In der Datenbank selbst war die Unterscheidung der unterschiedlichen Berufungskommissionen jedoch nicht vorgesehen, so dass sie sich nicht in den ursprünglichen Daten wiederfindet. Anhand der Titel der Kommissionen lassen diese sich jedoch weitestgehend identifizieren. Die Ergebnisse dieser Auswertung werden als weitere Attribute den Kommissionen zugeordnet.[21] Als Arbeitshypothesen liegen diese Daten in einem eigenen Graphen.[22]

Zur weiteren Charakterisierung wird eine systematische Übersicht über die Formen der Mitgliedschaften benötigt. In den Ursprungsdaten war jedoch hier lediglich ein Freitextfeld vorgesehen. Dieses ist jedoch für die weitere Auswertung ungeeignet. Daher wurden die verwendeten Begriffe in einem Vokabular zusammengefasst und Oberbegriffe definiert. Diese liegen als ein einfaches SKOS-Vokabular vor.[23]

8.3.3 Wikidata und weitere ergänzende Informationen aus der Datenbank

Die Personen in der Datenbank sind nicht systematisch mit externen Identifiern versehen worden. Wir holen dies hier zunächst automatisiert nach und nutzen dazu Daten aus *Wikidata*.[24] Wir hatten hierbei mit erheblichen Performanzproblemen zu kämpfen, wenn dies unter Einbeziehung der Abfrage externer SPARQL-Endpunkte erfolgen sollte [265] und benutzen daher eine Kopie des Wikidata-Datensatzes. Dazu nutzen wir das von *metaphacts* bereitgestellte Docker-Image [152], das wiederum auf von HOBBIT [111] bereitgestellte Daten zurückgeht. Die Identifikation von Einträgen der Personen-Datenbank und Wikidata erfolgt nach Namen und Geburtsdaten.[25] Eine händische Nachkorrektur ist jedoch noch notwendig und wurde bisher nur bei sehr wenigen Datensätzen durchgeführt.[26] Auswertungen, die sich auf Daten beziehen, die durch die Verknüpfung entstehen, sind daher mit der entsprechenden Vorsicht vorzunehmen. Trotzdem erhalten wir zusätzliche Informationen aus dieser Verknüpfung, beispielsweise einen direkten Verweis auf andere Identifier, wie die der *GND* oder *VIAF*.[27]

Abschließend ergänzen wir den Datenbestand noch um zusätzliche Informationen über die Personen aus der Personen-Datenbank, wie etwa Mitgliedschaften in Parteien. Diese fassen wir in einem eigenen Graphen zusammen, da es sich hier auch um Daten handelt, die dem Personenschutz unterliegen und dementsprechend nicht öffentlich zugänglich gemacht werden können. Dieser Graph wird daher in den zur Veröffentlichung vorgesehenen Daten nicht enthalten sein.

[21] Dazu wird eine eigene Unterklasse von **E55_Type** verwendet. Das Identifizierungsverfahren ist in einem Notebook *Add more details to the commission.ipynb* dokumentiert.

[22] *gmpgg:workingdataCommissions*

[23] *gmpgg:classify_commission_membership*

[24] Siehe Abschnitt 6.8.

[25] *Add_more_details_from_wikidata.ipynb*

[26] Stand: Januar 2018

[27] Auch dieses wird in 6.8 eingeführt.

Alle Daten fassen wir in einen Graphen *gmpgg:persons_all*[28] mit allen Informationen aus der Datenbank und Wikidata zusammen. Dieser Graph enthält über eine 1 Million Statements und 250.000 Instanzen von Klassen (Abbildung 8.5).

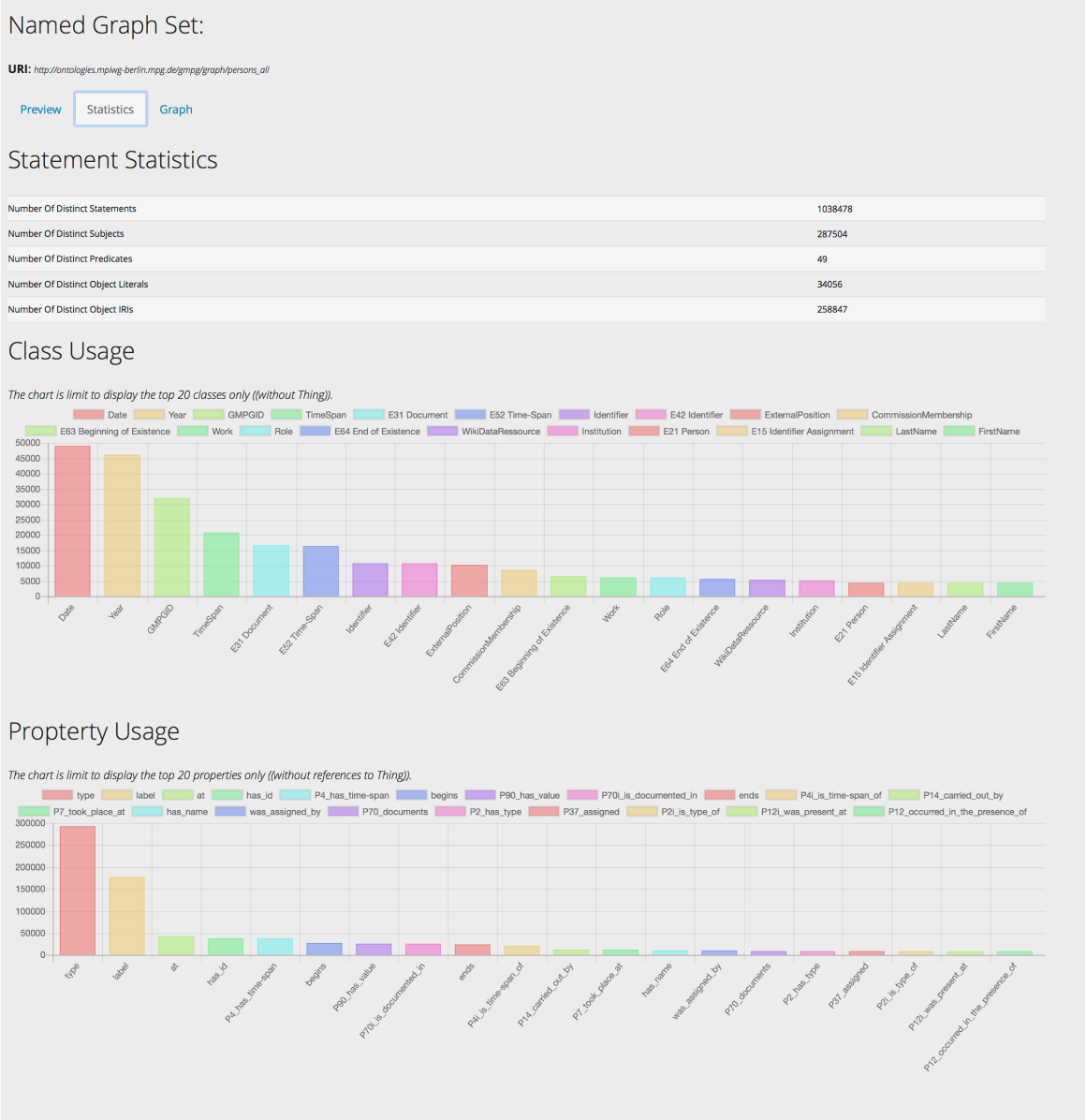


Abbildung 8.5: Übersicht alle Instanzen des ersten Imports

8.4 Multilevel-Struktur des Netzwerkes

Die nun im Triplestore vorhandenen Daten eröffnen die Möglichkeit einer ersten Darstellung des Graphen als Netzwerk. Das Netzwerk ist ein Multilevel-Netzwerk,[29] das aus mehreren bipartiten Teilnetzwerken besteht. Eine Gruppe von bipartiten Netzwerken setzt Personen über unterschiedliche andere

[28] Siehe Abbildung 12.3 in Kapitel 12.

[29] Wir greifen hier wieder auf die in Abschnitt 5.8 eingeführte Terminologie zurück.

Knoten miteinander in Beziehung. Zentral für die Fallstudie ist hierbei zunächst das bipartite Netz mit Personen und Kommissionen als Knoten. Die Kanten ergeben sich aus der Mitgliedschaft von Personen in den Kommissionen. Die Kommissionen sind selbst wiederum Teil bipartiter Netzwerke. Ein Typ von Knoten sind hier die Kommissionen. Superstrukturen, denen die Kommissionen angehören, sind der andere Typ. Diese Superstrukturen sind die Sektionen der MPG oder auch Einrichtungen wie der Wissenschaftliche Rat und der Senat der MPG.

Die semantischen Daten aus dem *Triplestore* erlauben die Konstruktion weiterer bipartiter Netzwerke. Diese sind zum Beispiel Personen, verbunden über verschiedene Institutionen, wie die Institute der MPG, aber auch über Parteien und andere akademische Institutionen. Zur Erstellung dieser Netzwerke können wir sowohl auf Daten aus dem Projekt als auch auf *Wikidata*, die *GND* und *VIAF* zurückgreifen. Die Struktur, die nun im Folgenden als Erstes betrachten werden wird, entspricht in der von Elisa Bellotti ([19, S.219] und Abschnitt 5.8) zusammengestellten Kategorisierung einem Typ, der mittels einer Modellierung im Rahmen *Multlilevel Network Analysis* MNA ([256]) untersucht werden kann.

8.4.1 Bipartite Teilgraphen – Kommissionen und Personen

Zur Vereinfachung betrachten wir zunächst den bipartiten Graphen der Kommissionen und Personen und berücksichtigen hier zusätzliche Superstrukturen bzw. andere mögliche bipartite Netzwerke zunächst lediglich als Attribute für die entsprechenden Knoten. Dies gilt insbesondere für die Zugehörigkeit der Personen zu Institutionen und der Kommissionen zu Sektionen sowie zu thematischen Clustern.[30] Die Ausweitung der Analyse zu einem *Multilevel-Netzwerk* mit Institutionen als dritte Ebene reißen wir nur kurz an.[31]

Wir versehen die Knoten mit Metadaten, die sich aus dem Triplestore abfragen lassen (Tabellen 8.1,8.2,8.3). Die dazugehörige SPARQL-Abfrage findet sich in Abbildung 12.4 im Anhang. Abbildungen der entsprechenden Netzwerke finden sich in den Abbildungen 8.6 und 8.7.

Atribut	Beschreibung	RDF-Property
label	Name der Person	sr:has_name/rdfs:label
id	ID in der Datenbank	sr:has_id

Tabelle 8.1: Attribute der Personen im Netzwerk

8.4.2 Reduktionen des Multilevel-Netzwerkes auf mono-modale Netzwerke

Ähnlich der Kooperationsbeziehungen der Fallstudie zur ART in Abschnitt 7.2 betrachten wir zunächst die Reduktion des Multilevel-Netzwerkes auf mono-modale Netzwerke. Diese sind ein Personennetzwerk (Abschnitt 8.5), wobei die Kanten hier durch gemeinsame Kommissionszugehörigkeit gebildet werden, sowie zwei verschiedene Versionen eines Netzwerkes der Kommissionen (Abschnitt

[30] Siehe Abschnitt 8.6.2.
[31] Siehe Abschnitt 8.9.

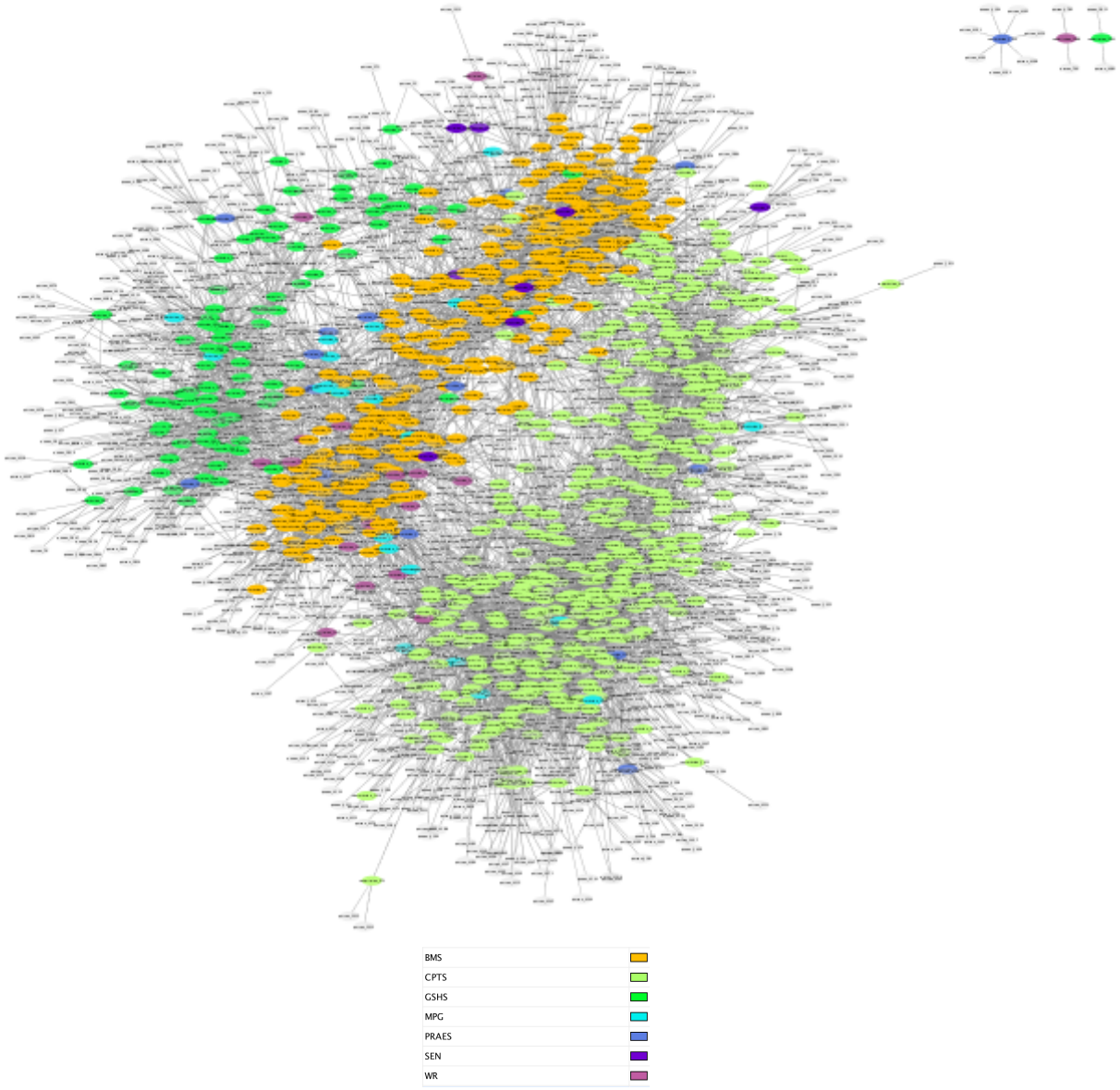


Abbildung 8.6: Personen in Beziehung zueinander über die Mitgliedschaft in Kommissionen: Grau = Personen; die Knoten in den andere Farben stellen Kommissionen dar.

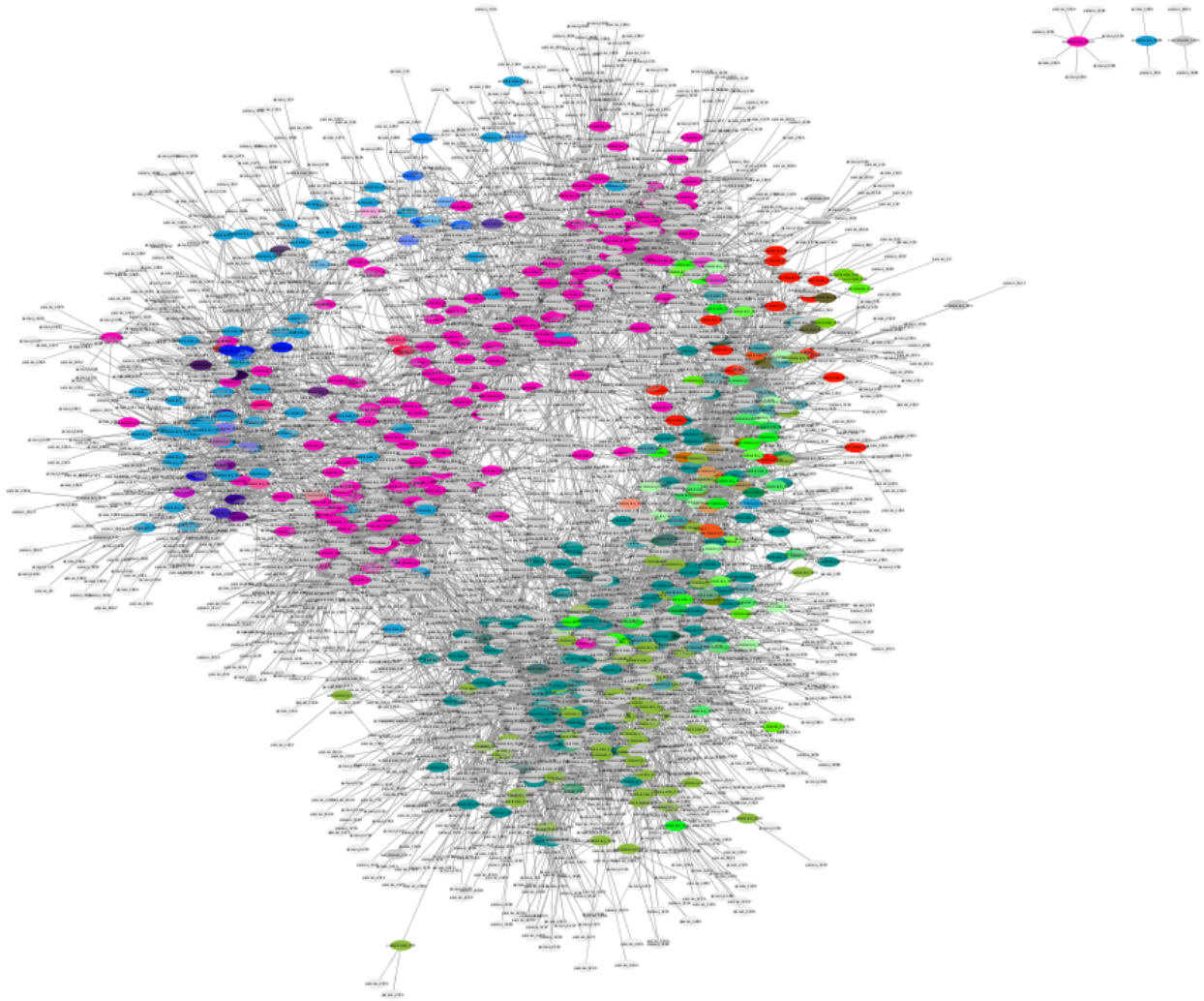


Abbildung 8.7: Personen in Beziehung zueinander über die Mitgliedschaft in Kommissionen: Grau = Personen, Knoten in anderen Farben geben thematische Cluster wieder.

Attribut	Beschreibung	RDF-Property
label	Name der Kommission	sr:has_name/rdfs:label
id	ID in der Datenbank	sr:has_id
sec	Sektion	sr:belongs_to
cls	Clusterzugehörigkeit	crm:P41i_was_classified_by/ crm:P42_assigned/rdfs:label
type1	Typ	crm:P2_has_type/rdfs:label
begin	Einsetzungsdatum	crm:P4_has_time_span/sr:begins (rdf:type = sr:Year)
end	Auflösung	crm:P4_has_time_span/sr:ends (rdf:type = sr:Year)

Tabelle 8.2: Attribute der Kommissionen im Netzwerk

Atribute	Beschreibung	RDF-Property
type	Typ laut Datenbank	crm:P2_has_type/rdfs:label
type2	Allgemeinere Kategorie des Types	crm:P2_has_type/skos:broader/rdfs:label
begin	Anfang	crm:P4_has_time_span/sr:begins (rdf:type = sr:Year)
end	Ende	crm:P4_has_time_span/sr:ends (rdf:type = sr:Year)

Tabelle 8.3: Attribute der Kanten im Netzwerk (Mitgliedschaft in einer Kommission)

8.6 und Abschnitt 8.8). Ziel dieser vergleichenden Untersuchung ist, Wege hin zu einer vollständigen Multilevel-Analyse im Sinne von Tom Snijders ([221]) finden. Gesucht sind dazu historische Argumente für die Bewertung des Einflusses der unterschiedlichen Typen der Bindungen innerhalb des komplexen Netzwerkes. Die Verbindung zwischen der Graphenstruktur im Triplestore und der Netzwerkanalyse ermöglicht hierbei ein flexibles Austesten verschiedener Netzwerkstrukturen auf der Grundlage einer nach inhaltlichen Gesichtspunkten wohldefinierten Datenbasis.

8.5 Personennetzwerke

Wir betrachten zunächst die Projektion[32] des oben beschriebenen bipartiten Netzwerkes auf die Personen als einziger Typ von Knoten. Gemeinsame Mitgliedschaft in Kommissionen begründet eine Verbindung zwischen den Personen (Abb. 8.9 und 8.8) .[33]

[32] In der Terminologie von Snijders ist dieses das erweiterte Netzwerk der Personen.
[33] Zur Projektion selbst gehört ein Notizbuch: *generate_graphs_commissions_all.ipynb*.

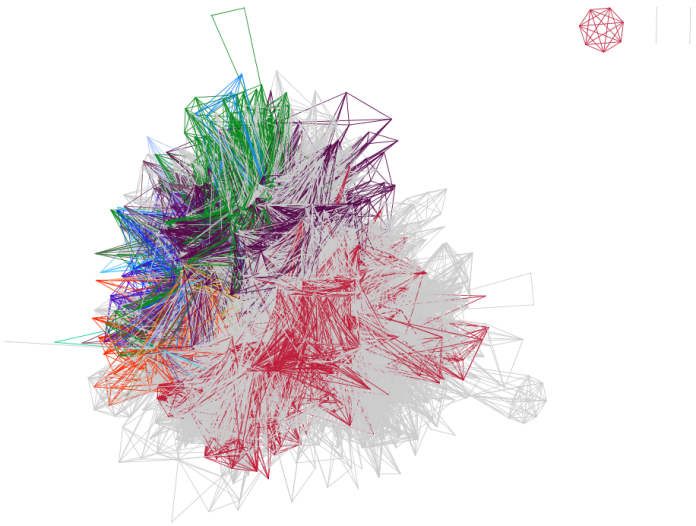


Abbildung 8.8: Person in Beziehung zueinander über die Mitgliedschaft in Kommissionen (Farben der Kanten gemäß der Clusterzugehörigket der Kommissionen, wie oben, grau kein Cluster)

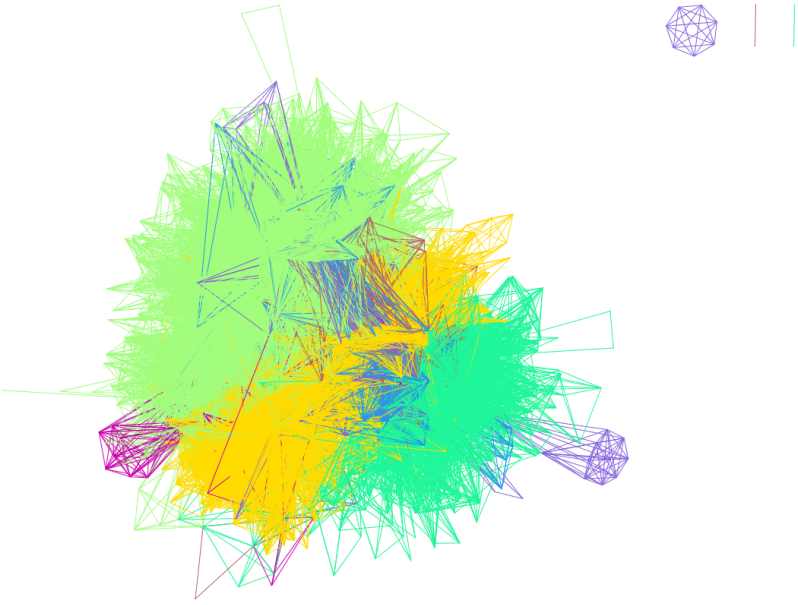


Abbildung 8.9: Person in Beziehung zueinander über die Mitgliedschaft in Kommissionen (Farben der Kanten gemäß der Sektionszugehörigkeit der Kommissionen)

8.5.1 Dynamische Entwicklung des Personennetzwerkes über die Zeit

Im Zentrum unserer Fragestellung steht wie in der vorhergehenden Fallstudie die dynamische Entwicklung des Netzwerkes. Auch hier stellt sich die Frage nach den Zeitintervallen, die für die Netzwerkstudien relevant sind. Analog zur Frage des Nachwirkung des Einflusses einer Kooperation unter Wissenschaftlern im Feld der ART, gilt es auch bei den Personennetzwerken dieser Fallstudie, Argumente dafür zu finden, wie lang die Zeiträume angesetzt werden sollten, über die Bindungen nachwirken. Hierbei sind wiederum historische Annahmen auf der einen Seite und Ergebnisse der Netzwerkanalyse auf der anderen Seite in Beziehung zu setzen. Annahmen über die Zeitintervalle aus historischer Sicht sind hierbei nicht zuletzt von der Einschätzung abhängig, wie mögliche Korrelationen zwischen den Präsidentschaften und Einflussstrukturen, Korrelationen zu gesellschaftlichen Entwicklungen oder auch Periodisierungen, die sich aus persönlichen Karrierewegen der betrachteten Personen ergeben, zu interpretieren sind. Eine der Fragen ist unmittelbar, ob die Amtsperioden der Sektionsvorsitzenden bzw. des Präsidenten von 3 bzw. 6 Jahren sich in den Strukturveränderungen der Netzwerke über die Zeit niederschlagen. In der vorhergehenden Studie hatten wir wenige historische Annahmen über die Zeitintervalle gemacht, die wir betrachten, und den Schwerpunkt vielmehr auf die Frage gelegt, ob die Variation des Nachhallparameters Veränderungen der Interpretation erzwingt und wie sich dieser auf die strukturellen Entwicklungen auswirkt. Erst im zweiten Schritt wurden dann die in den Netzwerken ablesbaren Entwicklungen auf historische Perioden bzw. Ereignisse bezogen. In dieser Fallstudie schauen wir jedoch stärker auf Entwicklungen, die sich durch extern angenommene Periodisierungen ergeben. Aus der Sicht einer Multilevelanalyse geht es hier letztendlich um das Problem der Wechselwirkung von externen Einflüssen mit dem Netzwerk. Hierbei können wir uns zwar von einzelnen Annahmen leiten lassen, die sich aus historischen Überlegungen ergeben, wie die unterschiedlichen Dynamiken innerhalb der verschiedenen Sektionen oder Vorstellungen über relevante Akteure. Welche Charakteristiken des Netzwerkes, die sich aus SNA und MLA ergeben, dafür jedoch aussagekräftig sind, bleibt zunächst offen. Visualisierungen helfen hier, einen ersten Eindruck von diesen Entwicklungen zu gewinnen.[34]

8.5.2 Jahresnetzwerke

Bereits die Frage, welche Kanten und Knoten des Netzwerkes in einem bestimmten Jahr zu berücksichtigen sind, wirft hierbei Probleme auf, die ohne Annahmen über die Arbeit der Kommissionen nicht zu lösen sind. Angaben für den Eintritt in eine Kommission, die durch Quellen belegt sind, finden wir bei 9302 von 9388 Mitgliedschaften, Enddaten jedoch nur für 6903. Wir sind daher auf Annahmen angewiesen und setzen in der Regel als Enddatum für die Mitgliedschaft in Kommissionen das Ende der Kommission selbst. Andere Fehlerquellen sind systematischer Natur. Es liegen uns nur Jahresangaben vor und keine genauen Daten. Damit impliziert die Angabe eines Jahres ein höchstwahrscheinlich

[34] Die dynamische Darstellung in Form einer Animation über den Zeitablauf erwies sich dabei als wenig hilfreich. Stattdessen wurde die Möglichkeit gesucht, zeitliche Schnitte visualisieren zu können und möglichst zu variieren. Siehe dazu Abschnitt 6.4.1.

verzerrtes Bild.[35] Wesentlicher jedoch ist auch hier die aus der Fragestellung abgeleitete Unsicherheit, wann der Einfluss von Bindungen sowohl im Hinblick auf Einzelpersonen als auch auf die MPG als Ganzes abnimmt. Die Zeitskalen und Annahmen über Gewichtungen von Kanten müssen hierbei in beiden Fällen nicht übereinstimmen. Daher muss auch in dieser Fallstudie wiederum intensiv der Frage nachgegangen werden, welche Auswirkungen die Zusammenfassung längerer Zeiträume auf die Strukturen des Netzwerkes hat. Dazu werden die Graphen für alle Jahre zunächst einzeln berechnet und jeweils als *GraphML-File* für alle Jahre gespeichert.[36]

8.5.3 Entwicklung des Netzwerkes

Um die Auswirkungen der Festlegung der Intervallgröße auf die Netzwerkcharakteristiken untersuchen zu können, bilden wir Verlaufskurven für eine Reihe von Charakteristiken[37] der Netzwerke über den gesamten Untersuchungszeitraum. Insbesonders betrachten wir hierbei die mit den Amtsperioden von 3 bzw. 6 Jahren verbundenen Intervalllängen. Wir betrachten jedoch bewusst jeweils die entsprechenden Zeiträume ausgehend von allen Jahren im Untersuchungszeitraum und nicht lediglich die Jahre der Amtswechsel, da wir mögliche Verzögerungseffekte nicht ausschließen können. Wir betrachten auch hier wie in der vorhergehenden Studie wieder vergleichende Graphen, die sowohl die Rolle von zentralen Personen im Netzwerk beleuchten helfen als auch Einblicke über die Entwicklung der Gesamtstruktur ermöglichen. Notizbücher[38] in Kombination mit *Cytoscape* und *Gephi* helfen hierbei, einen Überblick zu gewinnen. Abbildung 8.10 zeigt die Entwicklung der Graphen für ausgewählte Jahre mit Intervalllängen von 6 Jahren. Andere Intervalllängen geben qualitativ ein ähnliches Bild. Wir sehen über die gesamte Untersuchungsperiode hinweg, dass sich die Sektionen jeweils sehr deutlich von einander abgrenzen. Die Verbindungen zwischen den Sektionen entwickeln sich jedoch im Laufe der Zeit sehr unterschiedlich. Perioden, in denen es sehr viele Verbindungen zwischen den Sektionen gibt, wechseln sich mit Phasen, in denen nur wenige Kontakte bestehen, ab. Wir sehen weiterhin einen wachsenden Einfluss von intersektionellen Kommissionen (rot) für die Verbindung von Personen. Gleichzeitig fällt eine unterschiedliche innere Struktur der Sektionen auf, die auf unterschiedliche Formen der Kooperationen hinweisen.

[35] Im Extremfall fallen Kommissionen, die Anfang Januar beendet wurden, in die gleiche Kategorie wie Kommissionen, die im Dezember eines Jahres beginnen.

[36] Dazu dient *generate_graphs_commissions.ipynb*. Siehe auch Abschnitt 6.2.4 zum Umgang mit Jahresgraphen, sowie Abbildung 6.1 in Abschnitt 6.4.1.

[37] *doi:21.11103/dataverse.KMTEY5*/

[38] Dazu dienen die Notizbücher *Verlauf-Personen-Gesamt.ipynb* sowie *Clustering-Personen.ipynb*.

Abbildung 8.10: Mit Pythonnotizbüchern und Cytoscape erstellte Übersicht über die Entwicklung des Personennetzwerkes, Farben der Kanten geben Sektionszugehörigkeit der Kommissionen wieder. (Jeweils Zeiträume von 6 Jahren sind zusammengefasst.)

Wir vergleichen den Verlauf unterschiedlicher Charakteristiken über die Zeit (Abb. 8.11), um einen quantitativen Überblick über die Strukturveränderungen der Sektionen zu gewinnen.[39] Von Interesse ist das Verhältnis der Sektionen zueinander. Abbildung 8.12 zeigt hierbei jeweils den Quotienten der Charakteristiken (CPTS/BMS). Bei letzterem sehen wir deutlich, dass die BMS eine dichtere Netzwerkstruktur aufweist als die CPTS. Die Entwicklung des Radius aber auch des maximalen Degrees, geben erste Hinweise auf mögliche Periodisierungen in den Graphen. Wir sehen eine erste Phase bis etwa 1968, dann eine neue Phase von 1968–1989 und dann eine nach 1989.

[39] *Verlauf-Personen-Gesamt.ipynb*

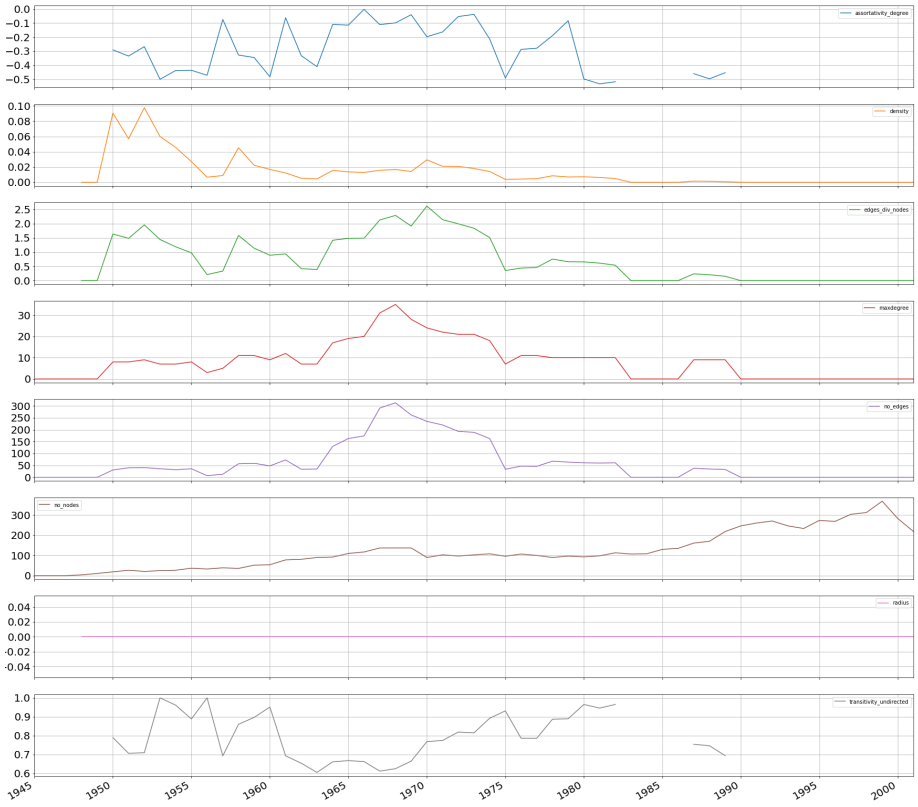


Abbildung 8.11: Zeitliche Entwicklung des Personenetzwerkes von 1945 - 2000

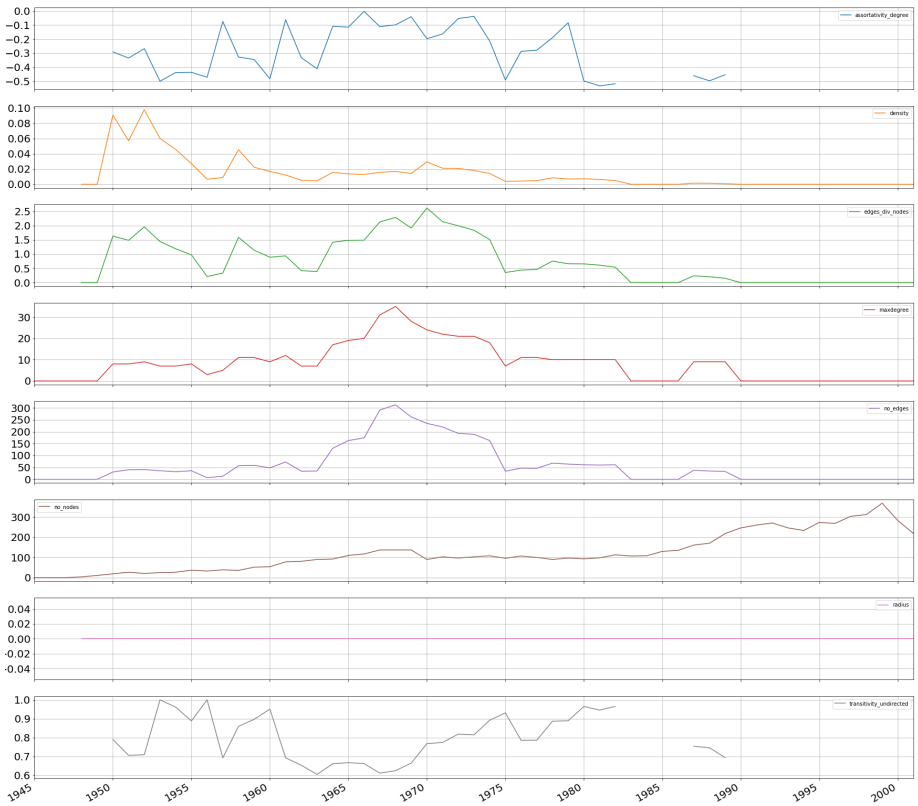


Abbildung 8.12: Zeitliche Entwicklung des Personenetzwerkes von 1945 - 2000 Verhältnis cpt und bms

Neben den Sektionen wurde vom Projekt eine Aufteilung der Arbeitsfelder in thematische Cluster vorgenommen. Ausführlicher kommen wir darauf noch in Abschnitt 8.6.2 bei der Diskussion der Kommissionen selbst zurück. An dieser Stelle sei daher nur die Bemerkung angebracht, dass auch die thematischen Cluster über weite Zeiträume in den Netzwerken deutlich erkennbar sind und als ein Kriterium zur Charakterisierung von Subgraphen herangezogen werden können (Abb. 8.13).

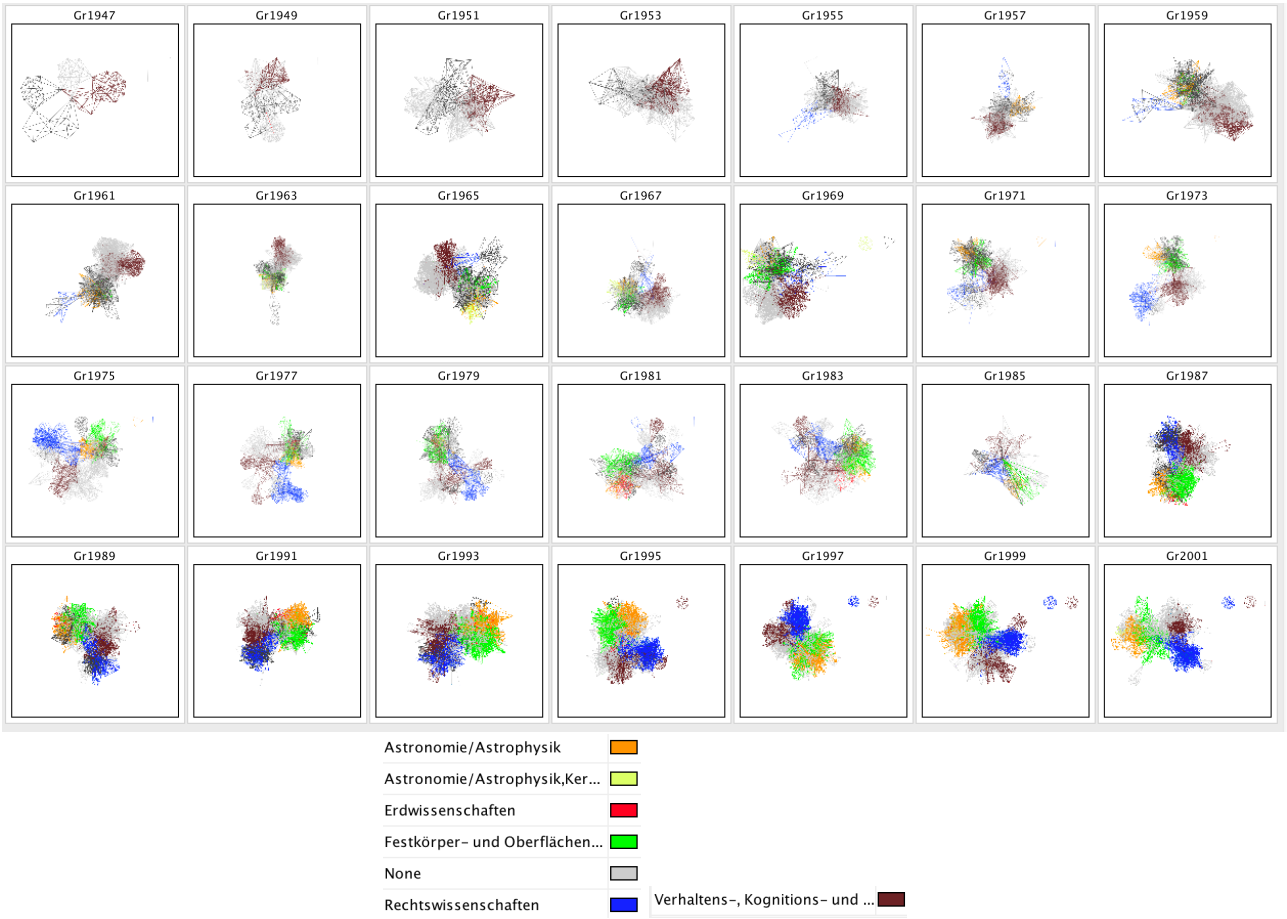


Abbildung 8.13: Zeitliche Entwicklung des Personennetzwerkes von 1945 - 2000, Thematische Cluster, Grundlage der Visualisierung sind jeweils Force-Modelle ohne Gewichtungen der Kanten. Es können aber mehrere Kanten zwischen einzelnen Knoten bestehen.

8.5.4 Rolle von Einzelpersonen

Schauen wir zunächst auf die Rolle von Einzelpersonen. Auch hier steht für die interaktive Analyse ein Notizbuch zur Verfügung.[40] Es erlaubt die Anzahl der Jahre, die zusammengefasst werden sollen, sowie der Anzahl der Personen pro Jahr, die angezeigt werden sollen. Zusätzlich erlaubt es, einen Parameter zu bestimmen, wie oft eine Person zu den ersten 20 in der Reihenfolge der höchsten Zentralität gehören muß, damit sie in den Übersichten berücksichtigt wird (Abb. 8.14 und 8.15).

[40] *Verlauf-Personen.ipynb*

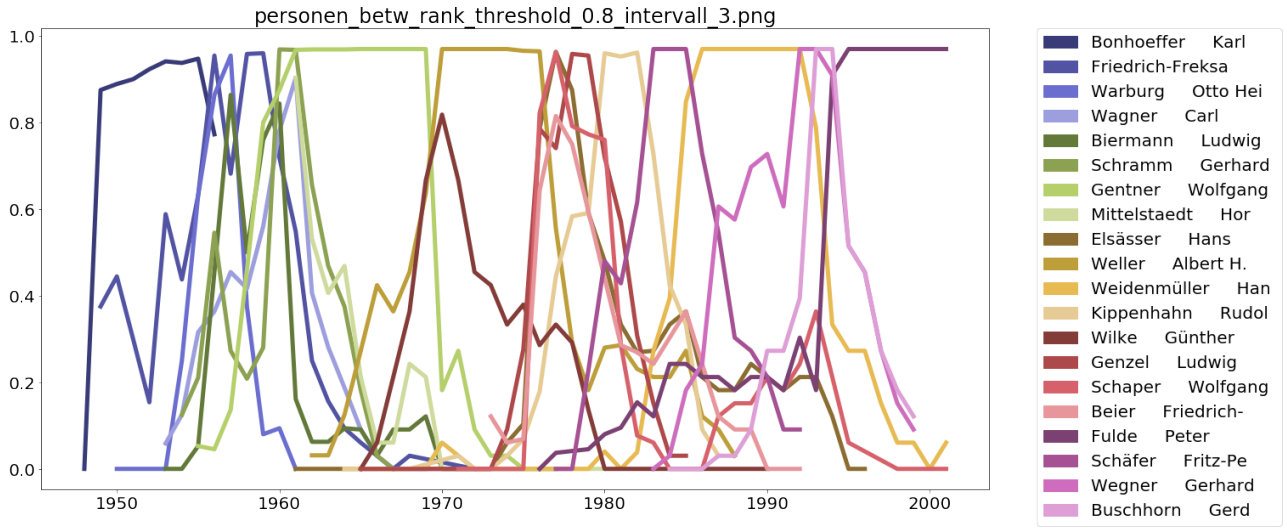


Abbildung 8.14: Personen, nach Rang (siehe 6.2.5) und Threshold 0.8 in den Top 100 bzgl. Betweenness-Zentralität auftauchen.

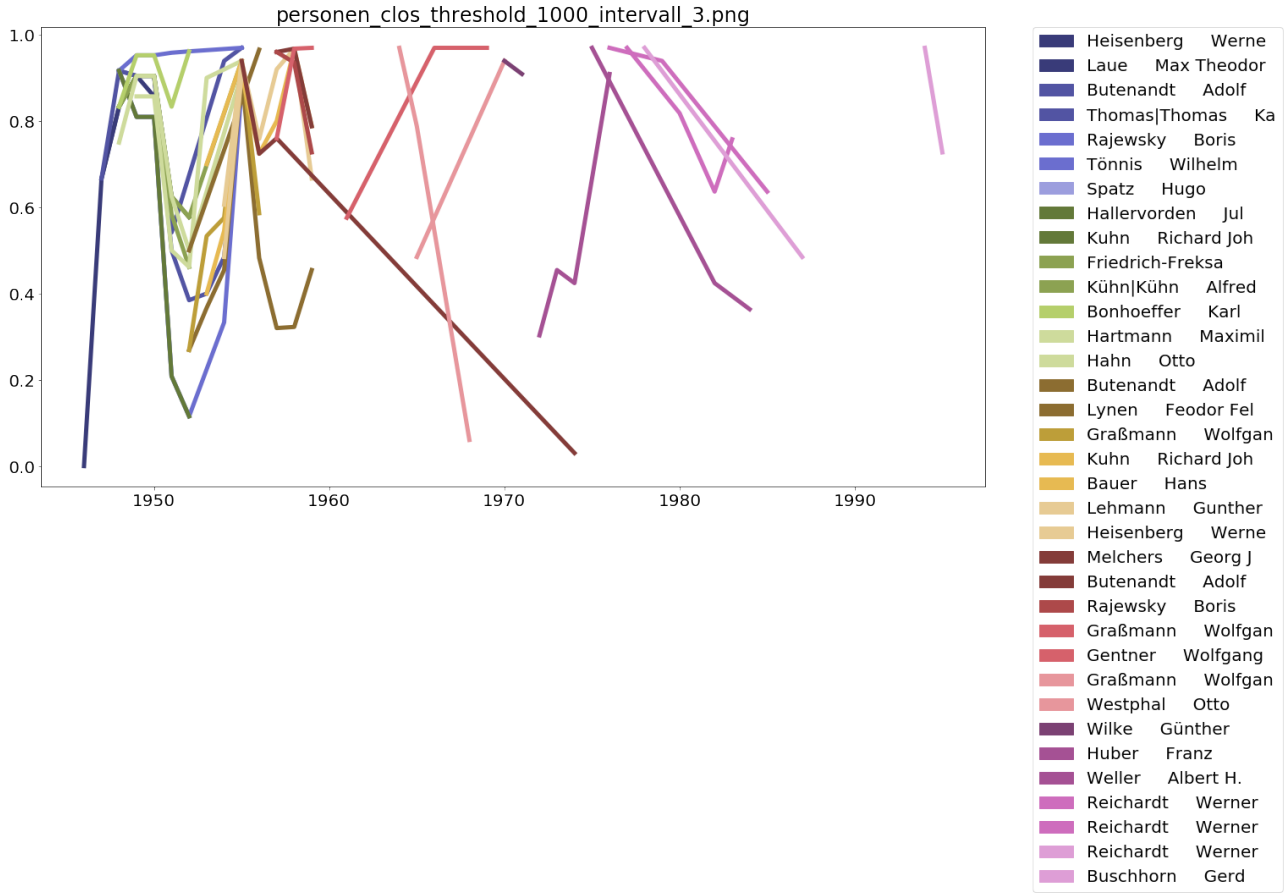


Abbildung 8.15: Personen, nach Rang (siehe 6.2.5) und Threshold 0.9 in den Top 100 bzgl. Closeness-Zentralität auftauchen.

Den Verlauf der Netzwerkentwicklung können wir mittels eines Notebooks[41] unter verschiedenen

[41] *Clustering-Personen.ipynb* und *Verlauf-Personen.ipynb*

Kriterien näher untersuchen. Nur wenige Personen haben die Funktion eines Mittlers zwischen den einzelnen Sektionen. Diese Entwicklung ist zunächst nicht überraschend, jedoch wird diese durch die quantitative Analyse nachweisbar, während sie aus qualitativer Sicht lediglich als plausible Annahme gemacht werden konnte (Abb. 8.16 und 8.17).

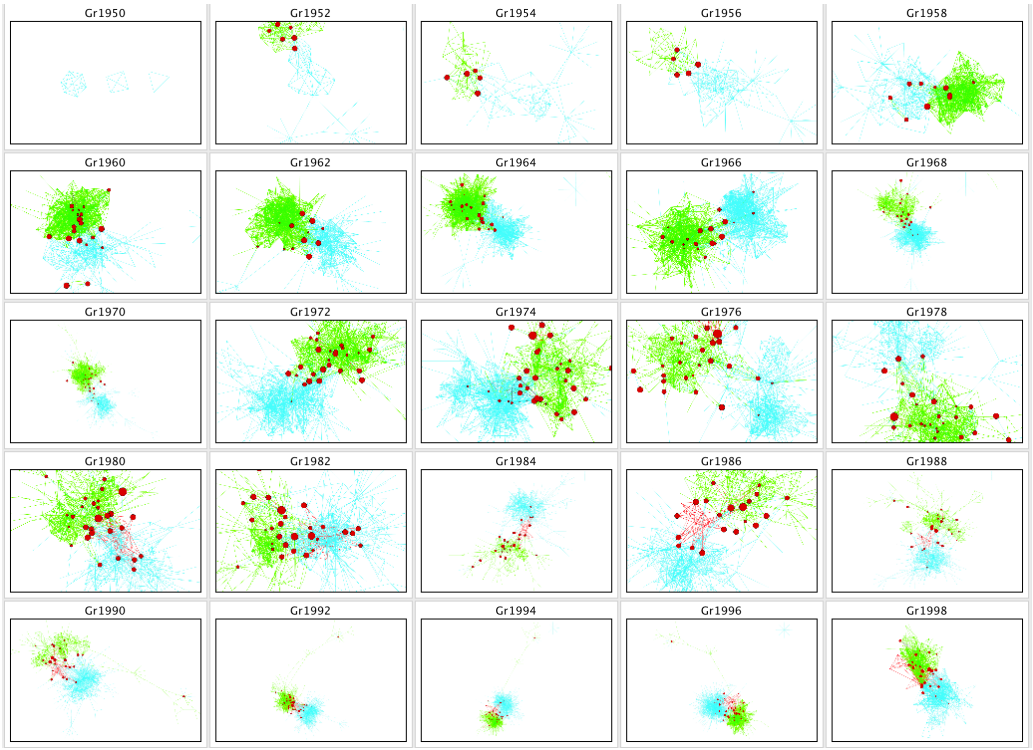


Abbildung 8.16: Mit Pythonnotizbüchern und Cytoscape erstellte Übersicht über die Entwicklung des Personennetzwerkes, Farben der Kanten geben Sektionszugehörigkeit der Kommissionen wieder. Die Größe der Knoten ist ein Maß für die Anzahl der intersektionellen Kontakte (Streuung der Verteilung).

8.5.5 Funktionsträger und ihre Rolle in dem Kommissionen

Die Beurteilung der Rolle von Personen, die Ex-officio-Mitglieder der Kommissionen sind, ist nicht offensichtlich und wirft grundsätzliche Fragen über die Aussagekraft der quantitativen Netzwerkanalysen auf. Unter dem Gesichtspunkt des Austausches von Wissen existieren sowohl Argumente dafür, diese in den Kommissionen zu vernachlässigen bzw. Verbindungen, die aufgrund dieses Personenkreises zustande kommen, niedriger zu bewerten, als auch für das Gegenteil. Die Entwicklung der unterschiedlichen Personenkreise im Hinblick auf ihre Stellung in den Netzwerken lässt sich mit den vorliegenden Daten jedoch zumindest quantitativ nachprüfen. Die Kombination Pythonnotizbüchern, Triplestore und Cytoscape ermöglicht es, auf einfache Art die Entwicklung der Stellung unterschiedlicher Personen innerhalb des Netzwerkes zu beurteilen. Als Beispiel zeigt Abbildung 8.18 die Stellung der Präsidenten über den Verlauf der Jahre[42] und Abbildung 8.19 die Stellung von Präsidenten und

[42] *Clustering-Presidents.ipynb*

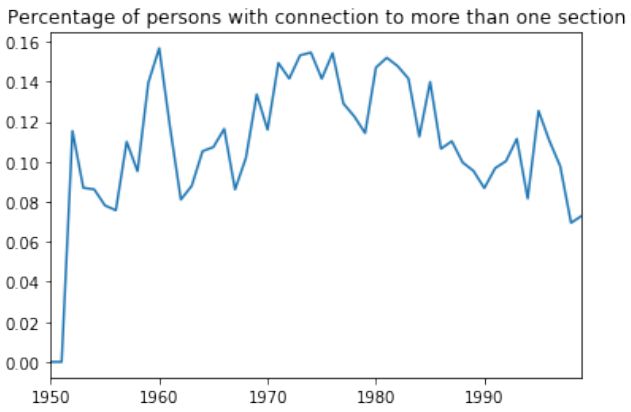


Abbildung 8.17: Prozentualer Anteil von Personen mit intersektionellen Kontakten.

Sektionsvorsitzenden.[43]

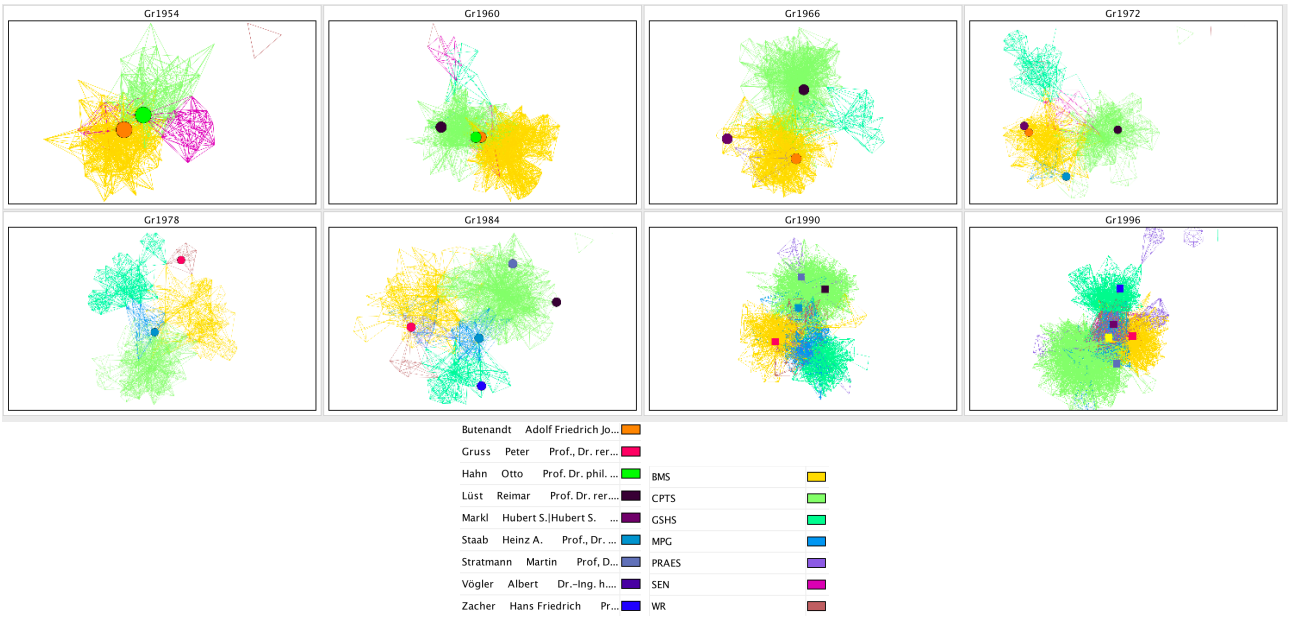


Abbildung 8.18: Mit Pythonnotizbüchern und Cytoscape erstellte Übersicht über die Entwicklung des Personennetzwerkes, Rolle der Präsidenten, jeweils Anfang einer Wahlperiode und 6 Folgejahre zusammengefasst.

[43] *Clustering-Presidents-Sektionsvorsitzende.ipynb*

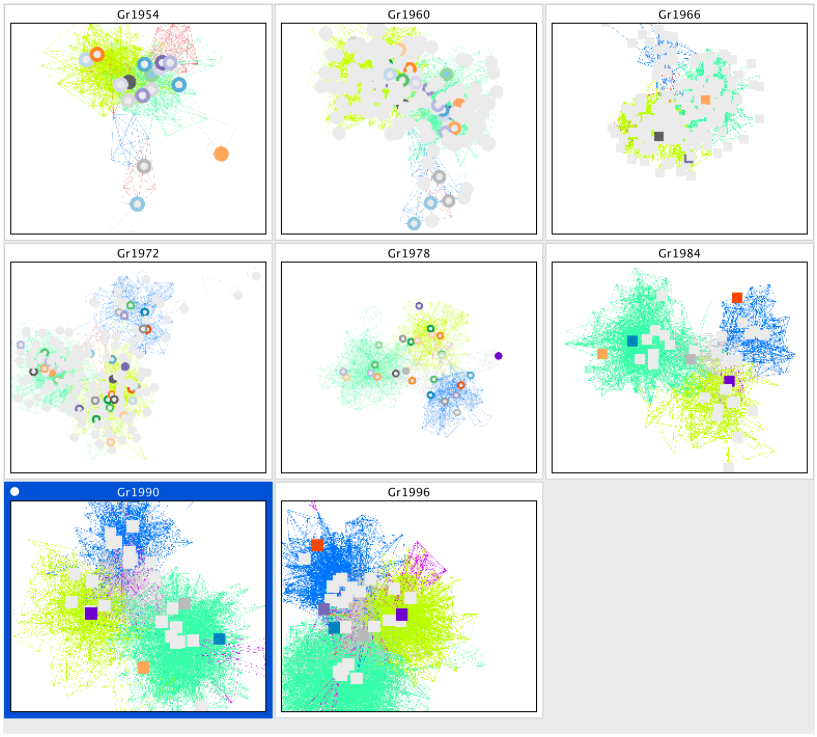


Abbildung 8.19: Mit Pythonnotizbüchern und Cytoscape erstellte Übersicht über die Entwicklung des Personennetzwerkes, Präsidenten erkennbar durch ausgefüllte Kreise, Sektionsvorsitzende durch Kreise mit grauem Mittelpunkt, falls diese nicht Präsidenten wurden, farbig in anderem Falle. Die Farben der Kanten stehen für die unterschiedlichen Sektionen, die Jahre sind jeweils Anfänge neuer Wahlperioden des Präsidenten. Es werden wieder jeweils 6 Jahre zusammengefasst, also jeweils die gesamte Wahlperiode.

Wir schauen uns genauer den Verlauf der Zentralitäten sowohl mittels eines Rankingverfahrens[44] als auch die Entwicklung der absoluten Zahlen an.[45]

Die Präsidenten

Die Stellung der Präsidenten in unseren Kommissionsnetzwerken ist eindeutig ablesbar, abgesehen von der Sonderrolle Otto Hahns in der Anfangsphase der MPG (Abb. 8.20). Mit Beginn ihrer Amtszeit sind diese faktisch nicht mehr präsent in unseren Daten. Die Präsidenten Zacher, als erster Präsident aus der Geistes-, Sozial- und Humanwissenschaftlichen Sektion(GSHS), und Hubert Markl, der von außen in die MPG kommt, kommen in den Zahlen gar nicht vor und Reimar Lüst nur marginal. Der Einfluss der Präsidenten ist daher mit diesen Methoden nicht messbar. Die Frage für die nächsten Schritte ist hier, wie dieses in den Entwicklungen der Netzwerke zu berücksichtigen ist. Insbesondere bei den Präsidenten Staab und Butenandt ist zu überlegen, ob wir ihre Rolle in den Netzwerken über ihr Ausscheiden aus den Kommissionen hinaus verlängern sollten. Ein Hinweis ist die Closeness. Sie verändert sich auffälligerweise nicht so stark wie die Betweenness (Abb. 8.21).

[44] Siehe Abschnitt 6.2.5.

[45] Die entsprechenden Analysen können wieder in einem Notizbuch *Clustering-Presidents-Sektionsvorsitzende.ipynb* nachvollzogen werden.

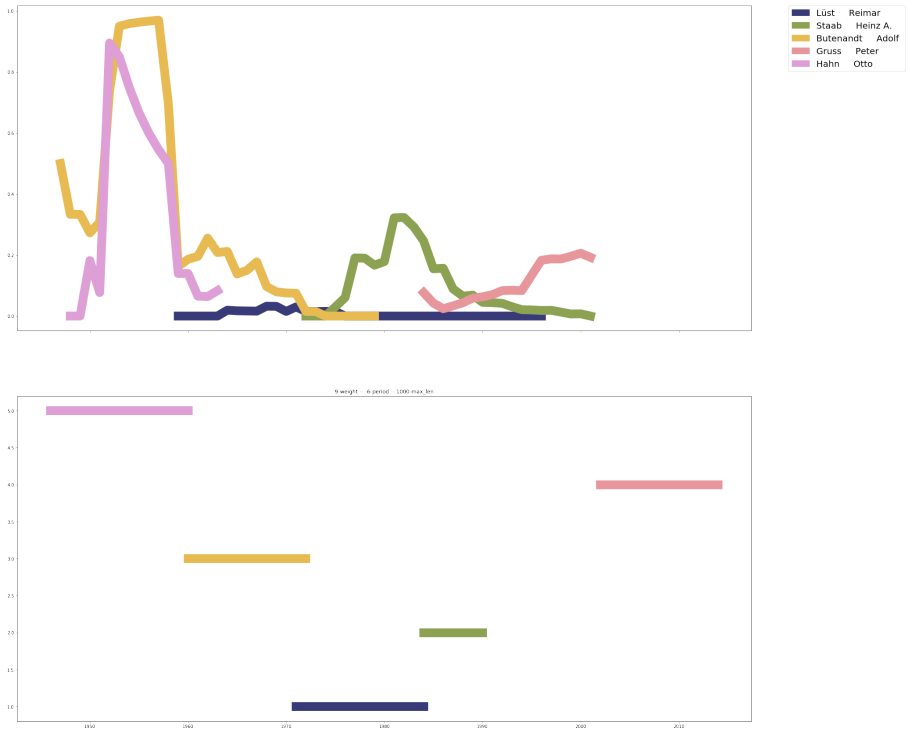


Abbildung 8.20: Entwicklung der Betweenness-Zentralität der MPG-Präsidenten (rang) - 6 Jahre, unteres Diagramm zeigt die Amtszeiten.

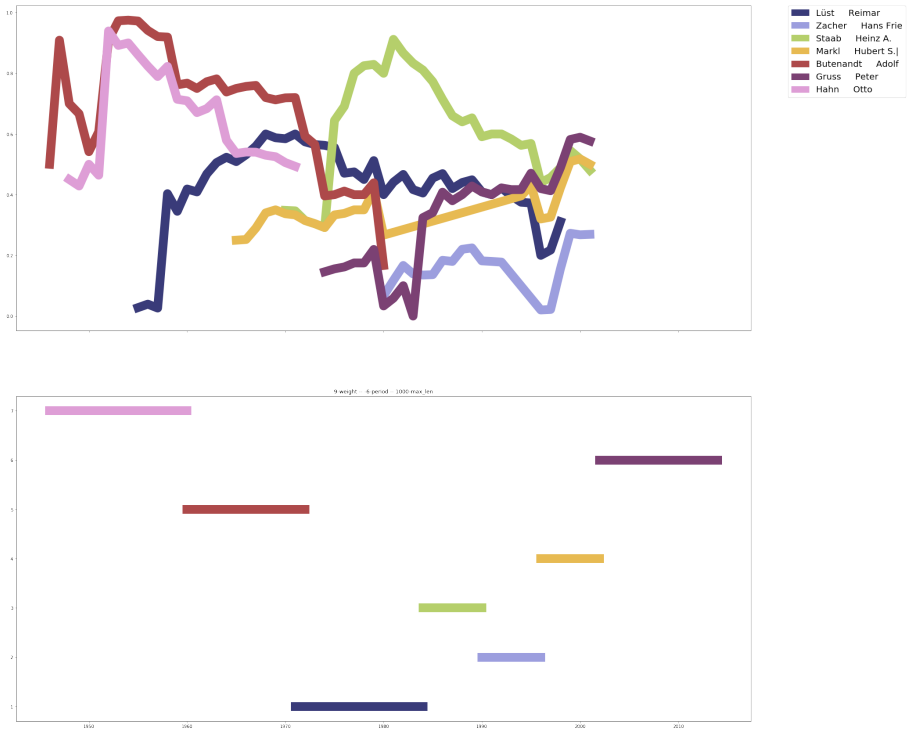


Abbildung 8.21: Entwicklung der Closeness-Zentralität der MPG-Präsidenten (rang) - 6 Jahre, unteres Diagramm zeigt die Amtszeiten.

Die Vizepräsidenten

Im Falle der Vizepräsidenten, die zugleich wissenschaftliche Mitglieder sind, ist die Entwicklung ähnlich wie bei den Präsidenten. Auch hier sehen wir bis auf zwei Ausnahmen, Konrad Zweigert und Herbert Walther, einen Abfall der Bedeutung in unseren Netzwerkdaten. Zweigert ist hierbei aus der GSHS, die generell eine Sonderrolle einnimmt, da sie wesentlich kleiner als die anderen Sektionen ist, und Walther hat eine der geringsten absoluten Werte, so dass es schnell zu statistisch bedingten Abweichungen kommen kann. Die Abbildung 8.22 zeigt hier die Entwicklung von Beginn ihrer Amtszeit im Vergleich 4 Jahre danach, wobei immer die zurückliegenden 4 Jahre berücksichtigt wurden (jeweils $\frac{\text{Anfangswert}-\text{Endwert}}{\text{Mittelwert}}$).

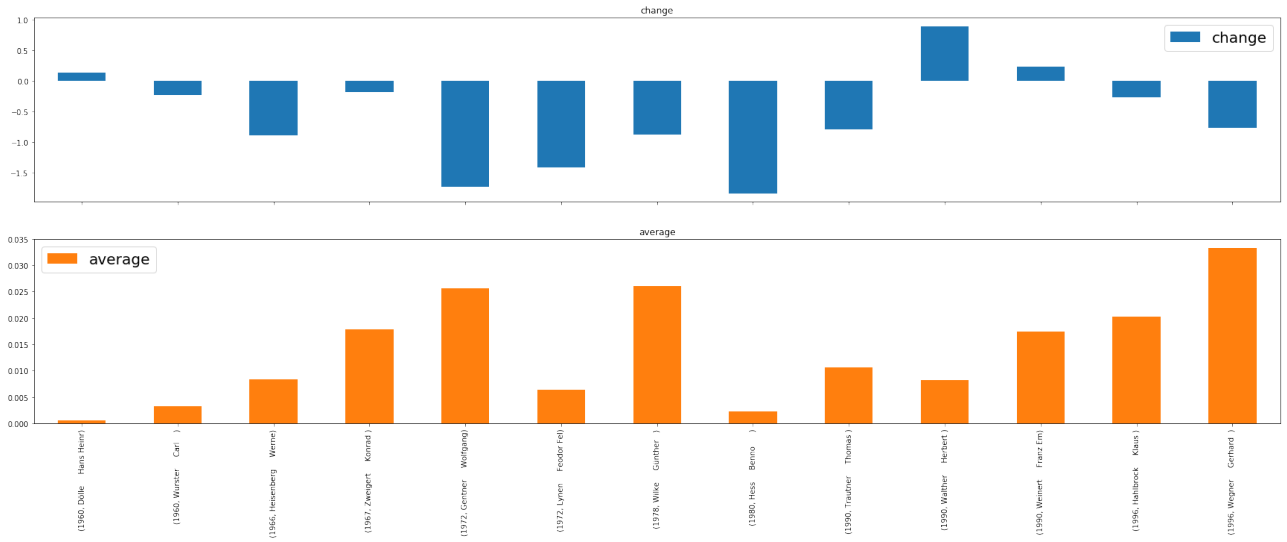


Abbildung 8.22: Normierte Mittelwerte der Betweenness der Vizepräsidenten (Mittelwert aus Beginn der Amtszeit, 4 Jahre nach Beginn) und rel. Veränderung , Beginn der Amtszeit, 4 Jahre nach Beginn, 4 Jahre zusammengefasst.

Sektionsvorsitzende

Auch bei den Sektionsvorsitzenden sehen wir ein ähnliches Bild (Abb. 8.23). Während ihrer Amtszeit sind diese als Funktionsträger weniger in den Netzwerken vertreten (im Sinne von Betweenness Centrality) als vorher. Zugleich sind in der Regel Sektionsvorsitzende auch vor ihrer Wahl relativ aktiv in den Sektionen. Wie sehr sich die Rolle der Sektionsvorsitzenden auswirkt, sehen wir später auch noch einmal in der Untersuchung der Kommissionsnetzwerke.[46]

Zusammenfassung: Rolle der Ex-Officio Mitglieder und weitere Schritte

Die Rolle der Mandatsträger in den Netzwerken aus den gegebenen Daten einzuschätzen ist problematisch. Entgegen unserer ersten Überzeugung, dass Mandatsträger – insbesondere Sektionsvorsitzende – das Netzwerk durch Überrepräsentanz aufgrund ihrer Ex-Officio-Anwesenheit verzerren,

[46] Siehe Abschnitt 8.7.

Abbildung 8.23: Normierte Mittelwerte der Betweenness der Sektionsvorsitzenden (Mittelwert aus Beginn der Amtszeit, Ende der Amtszeit) und rel. Veränderung , 2 Jahre zusammengefasst.

scheinen die Daten den umgekehrten Effekt zu zeigen.Überraschend gilt dieses auch für die Sektionsvorsitzenden. Diese erwerben sich ihre Stellung im Netzwerk in der Phase vor ihrem Vorsitz und behalten während dieser Zeit im Mittel ihren Status. Für Sektionsvorsitzende und Vizepräsidenten verhalten sich Closeness und Betweenness etwa parallel zueinander. Die Daten dazu lassen sich leicht mit dem zitierten Notizbuch nachvollziehen. Jedoch ist dies bei den Präsidenten anders – sie verlieren ihre Stellung im Netzwerk im Hinblick auf die Betweenness, für die Closeness ist dieser Effekt jedoch nicht so deutlich. Letzteres ist zugleich auch ein Artefakt der Problematik der Closeness in Fällen geringer Degrees und Betweenness. Einige wenige Kommissionen können die Stellung eines Mitgliedes deutlich verändern. In der Tat sind es lediglich zwei Kommissionen, denen Hans Zacher angehörte, diese jedoch haben einen hohen Vernetzungseffekt. Das wesentliche Problem unserer Quellen ist jedoch, dass die Protokolle Funktionsträger schlechter repräsentieren. In den Protokollen der von uns ausgewerteten Gremien werden diese als Mitglieder einer Kommission nur ausnahmsweise erwähnt, da hier in aller Regel nur die Aufnahmen von Mitgliedern explizit dokumentiert werden, die nicht ex officio an den Kommissionen teilnehmen. Hier sind weitere Fallstudien notwendig. So untersuchen wir zurzeit die Dokumente auch im Hinblick auf Worthäufigkeiten und impliziter Nennung von Personen, z. B. über ihre Titel.

8.6 Das komplementäre Netzwerk der Kommissionen

Dem Personennetzwerk stellen wir im Weiteren das Netzwerk der Kommissionen selbst gegenüber. Welche Bedeutung hat dieses Netzwerk? Der Leitgedanke ist dabei, dass über die Kommissionen neue Forschungsfelder erschlossen werden und sich neue Themenbereiche formieren. Unsere Hypothese hierbei ist, dass die Kommissionen zentrale Orte sind, in denen neue Themenbereiche diskutiert werden und schließlich in Form von Neuberufungen und Institutsgründen materialisiert werden. Wir

vermuten, dass über die Teilnahme an Kommissionen Ideen im Hinblick auf neue Felder transportiert werden. In diesen Netzwerken spielt die Bewertung unterschiedlicher Funktionen der Akteure in den Kommissionen unter Umständen eine wesentliche Rolle für die Gewichtung der Kanten. Wir suchen daher unter anderem nach Kriterien, die auf eine aktivere Rolle für die Gesamtpolitik der MPG und auf Einfluss in einzelnen Themenfeldern schließen lassen. Es existiert außerdem das oben[47] erwähnte Problem der Rolle der Amtsträger der MPG. Wir betrachten zunächst die dynamische Entwicklung des Kommissionsnetzwerkes. Wir verfolgen auch hier wie beim Vorgehen bei der Analyse der Personennetzwerke zunächst einen heuristischen Ansatz. Gibt es Auffälligkeiten, die sich schon aus der Visualisierung der Netzwerke ergeben? Zunächst schauen wir wieder auf die jährliche Entwicklung. Noch wesentlicher als bei den Personennetzwerken ist hier die Frage, wie lange Verbindungen noch über das Ende einer Kommission hinaus bestehen, das heißt, für welche Zeiträume wir annehmen, dass die Mitgliedschaft in einer Kommission Entscheidungen in künftigen Kommissionen beeinflussen. Eine klare Hypothese dazu konnten wir bisher nicht entwickeln, so dass wir zunächst pragmatisch der Frage nachgehen, welche Relevanz ein solcher freier Parameter für die Entwicklung des Netzwerkes und seiner Charakteristiken hat. Dazu vergleichen wir zwei unterschiedliche Netze. Das ist einerseits das Kooperationsnetzwerk, das dadurch entsteht, dass wir Kommissionen immer dann miteinander verbinden, wenn sie unmittelbar gemeinsame Mitglieder haben. Andererseits untersuchen wir ein Einflussnetzwerk, das dadurch entsteht, dass wir Kommissionen in Beziehung setzen, wenn in der Zukunft Mitglieder aus diesen Kommissionen gemeinsam in einer Kommission sitzen. In Analogie zum Kozitationsnetzwerk der vorhergehenden Fallstudie nennen wir dieses Netzwerk Koinfluenznetzwerk. Wir verbinden das erstere Netzwerk mit der Frage nach der Auswirkung von unmittelbarer Kooperation, das zweite mit der Frage, welche Themenfelder zur Formierung von neuen Feldern geführt haben. Fragen, die sich in beiden Netzwerken ergeben, sind die nach der Kontinuität über mehrere Forschergenerationen hinweg und die nach der Konvergenz von Themen. Welche Rolle spielen hier einzelne herausragende Personen, welche Institutionen und Forschungsfelder? Die Datenlage gibt zurzeit nur unvollständig Auskunft über diese Fragestellung. Vor allem die Zuordnung von Forschungsfeldern zu Personen ist bisher nur eingeschränkt möglich. Trotzdem zeigt sich, dass der in dieser Arbeit dargestellte Ansatz zumindest für die Hypothesenbildung vielversprechend ist. Im Folgenden soll ein Eindruck darüber gegeben werden, welche Schritte in diesem Kontext schon gegangen worden sind und wo sich aus unterschiedlichen Gründen noch Desiderate ergeben. Auch für diesen Abschnitt gilt, dass im Zentrum dieser Arbeit das methodische Handwerkszeug für die historische Arbeit mit Netzwerkdaten steht und die historische Interpretation der Ergebnisse zweitrangig ist. Die Frage, die im Raume steht, ist zunächst, ob dieser Ansatz überhaupt historisch zu rechtfertigen und an die im Felde arbeitenden Wissenschaftlerinnen und Wissenschaftler zu vermitteln ist. Wir haben uns entschlossen, die Kommissionen auszuschließen, die sich mit der Berufung von auswärtigen wissenschaftlichen Mitgliedern (AWM) beschäftigen.[48] Außerdem betrachten wir parallel zu allen Kommissionen jeweils noch das Netzwerk ohne die Kanten, die sich aus der Ex-Officio-Mitgliedschaft von einzelnen Personen ergeben.

[47] Siehe Abschnitt 8.5.5.

[48] Siehe dazu 8.3.2.

8.6.1 Verhältnis der Sektionen

Ausgangspunkt ist erneut der bipartite Graph aus Abschnitt 8.4, der sich aus den Personen und Kommission zusammensetzt. Dieses Mal projizieren wir den Graphen auf das mono-modale Netzwerk der Kommissionen.[49]

In allen Darstellungen werden jeweils die Knoten eingefärbt; die Farbe der Kanten ergibt sich dann aus den Farben bzw. Farbübergängen zwischen den Knoten. Dieses bedeutet, dass die Dichte der Farben in der graphischen Darstellung einen ersten Eindruck über die Zusammenarbeit zwischen Kommissionen nach den jeweilig gewählten und eingefärbten Attributen gibt.

Wir sehen einige Auffälligkeiten in Abb. 8.24. Die Sektionen sind deutlich zu unterscheiden. Während jedoch BMS und CPTS relativ eng verflochten sind, gilt dies nur teilweise für die GSHS, sie scheint enger mit der BMS verflochten als mit der CPTS. Dies ist jedoch teilweise ein Problem der zweidimensionalen Projektion. Es gibt einzelne Kommissionen, die eine relativ große Anzahl von Verbindungen zwischen GSHS und BMS haben, die in dieser Darstellung nicht hervortreten, dieses sind insbesondere Kommissionen aus dem Bereich der Linguistik. Einige Kommission aus der GSHS fallen deutlich heraus, als einzige einer Sektion zugeordnete Kommissionen liegen sie „mitten in der falschen Farbe'". Dieses ist die Intersektionelle Kommission „Musik-Institut" / MPI für Musik (Kommission 728) und die Berufung von Hans-Jörg Rheinberger an das MPI für Wissenschaftsgeschichte.[50]

Die optische Einordnung zeigt, wie vermutet, drei Schwerpunkte, die durch die Sektionen gebildet werden. Diese drei Zentren machen die 2D-Darstellung problematisch, da wir 3D-Effekte erwarten müssen. Auffällig sind außerdem noch einzelne Kommissionen aus dem Bereich der CPTS, die sehr stark in den Bereich der GSHS hineinragen und umgekehrt. Schauen wir etwas genauer in den Mittelbereich, so werden die Gemeinsamkeiten dort deutlich, wenn wir die Kommissionen hervorheben, in denen „biophysik", „medizin" und „linguistik" im Titel vorkommen.

[49] Ein direkter Zugriff ist über */netzwerke-der-mpg/kommissionen-und-kommissionen/kommissionen-kommissionen-verbunden-uber-gemeinsame-mitgliedschaften* möglich. Die entsprechenden Netzwerke werden, wie auch die Personennetzwerke in Abschnitt 8.5 mittels des Notizbuches *generate_graphs_commissions_all.ipynb* erzeugt und finden sich innerhalb unseres Dataverses unter *doi:21.11103/dataverse.YJL9C0*/. Entsprechend wird das Netzwerk ohne Ex-officio-Mitglieder erzeugt, auch dieses Netzwerk können wir direkt aus dem Triplestore mit SPARQL-Abfragen erzeugen. Das entsprechende Vokabular für die Form der Mitgliedschaft enthält die entsprechenden Information, so dass wir die Abfrage nur um einen entsprechenden Filter: filter (!regex(str(?ns_nm__type2_memb),"EX"))) erweitern müssen. Die Ergebnisse liegen in *doi:21.11103/dataverse.6CXXC1*/

[50] Die der BMS zugeordneten in Grün gekennzeichneten Kommissionen, die scheinbar am linken Rand in die GSGS hineinragen, sind ein Artefakt der 2D-Darstellung des Netzwerkes. Diese haben keine Beziehungen zur GSHS, sondern liegen dort aufgrund ihres Abstandes zur BMS und CPTS. Außerdem gibt es noch eine „grüne" Kommission inmitten der CPTS. Dies ist eine Senatskommission, die der BMS fehlerhaft zugeordnet wurde. Zwei weitere Kommissionen der GSHS, die zwischen der BMS und CPTS im Diagramm angesiedelt sind, sind einmal eine Senatskommission, und die andere Kommission (1674) hat ein sehr aktives Mitglied aus der CPTS, das eine große Anzahl von Kanten zu anderen Kommissionen aus der CPTS erzeugt. Dieses eine Mitglied hebt die Wirkung der anderen Mitglieder der Kommission, die alle der GSHS angehören, in dieser Darstellung auf.

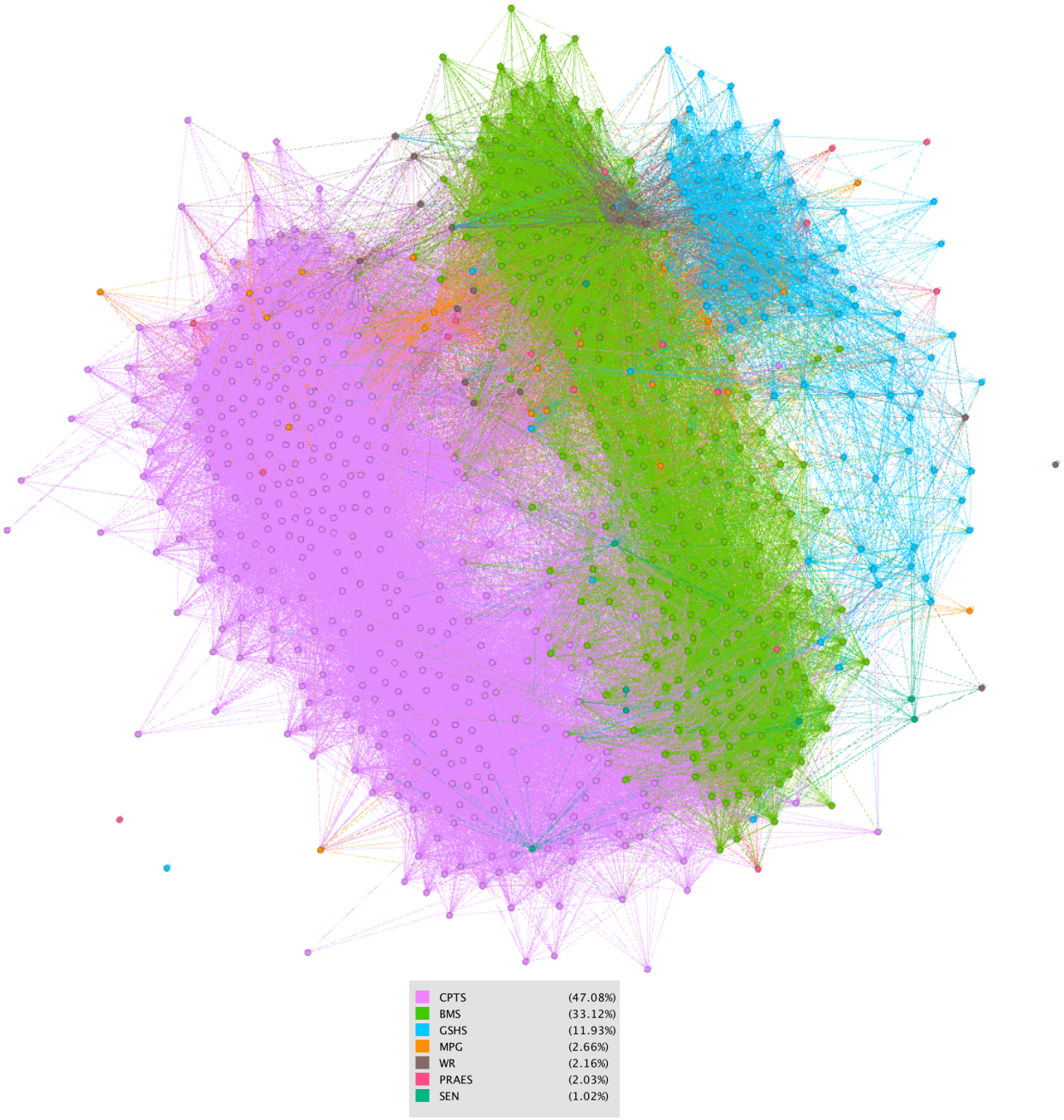


Abbildung 8.24: Kommissionen (ohne AWM) verbunden über gemeinsame Mitglieder, Einfärbung nach Sektionen, Darstellung mit Gephi, Fruchtermann-Reingold (Area 10000, Gravity 10.0, Speed 1)

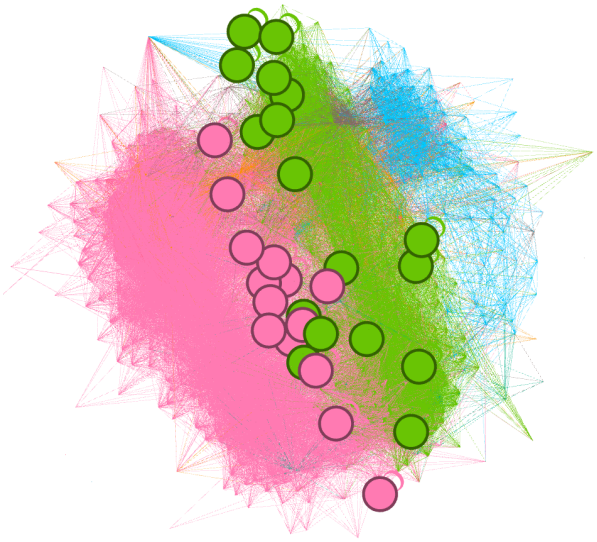


Abbildung 8.25: Biophysik und Medizin, als größere Punkte dargestellt, bilden die Grenzflächen.

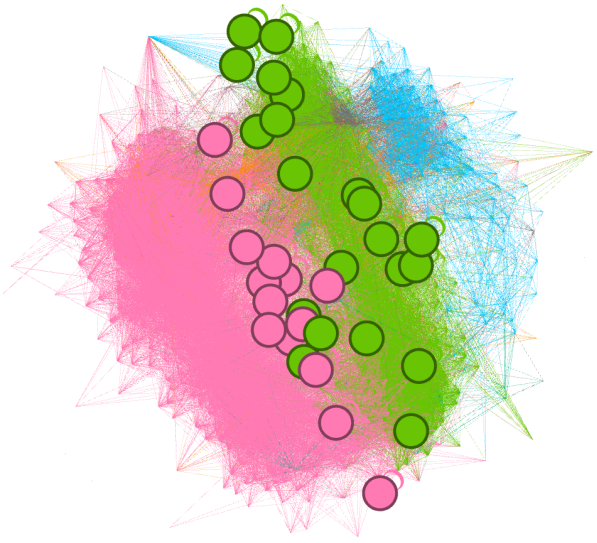


Abbildung 8.26: Linguistik, Biophysik, Medizin

Man sieht in den Abbildungen 8.26 und 8.25 deutlich die Bedeutung der biophysikalischen und medizinischen Forschung für die Verbindung der Sektionen. Diese sind alle an den Grenzbereichen der Sektionen aufzufinden.

8.6.2 Thematische Cluster

Bevor wir weiter auf Interpretationsmöglichkeiten eingehen und insbesondere auf die Analyse der zeitlichen Entwicklung zurückkommen werden, nehmen wir zunächst noch eine weitere Analyseebene hinzu. Den Kommissionen wurden Informationen über mögliche thematische Cluster zugeordnet. Die Zuordnung selbst erfolgte manuell und ist ebenfalls in der Personendatenbank verzeichnet. Für die Cluster selbst existiert ein kontrolliertes Vokabular, das zunächst in ein SKOS-Vokabular überführt wird.[51] Die Klassifikation der Kommissionen selbst geschieht über die Standard CRM-Klasse **E17_Type Assignment**. Da diese Typisierung sehr subjektiv nach Einschätzung der Wissenschaftlerinnen des Projektes vorgenommen wurde, versehen wir nicht die Kommissionen selbst direkt mit dem Typ, sondern wählen den Zwischenschritt über ein Assignment, da wir dann in der Lage sind, die Zuweisung selbst mit Kommentaren zu versehen (Abbildung 8.27).

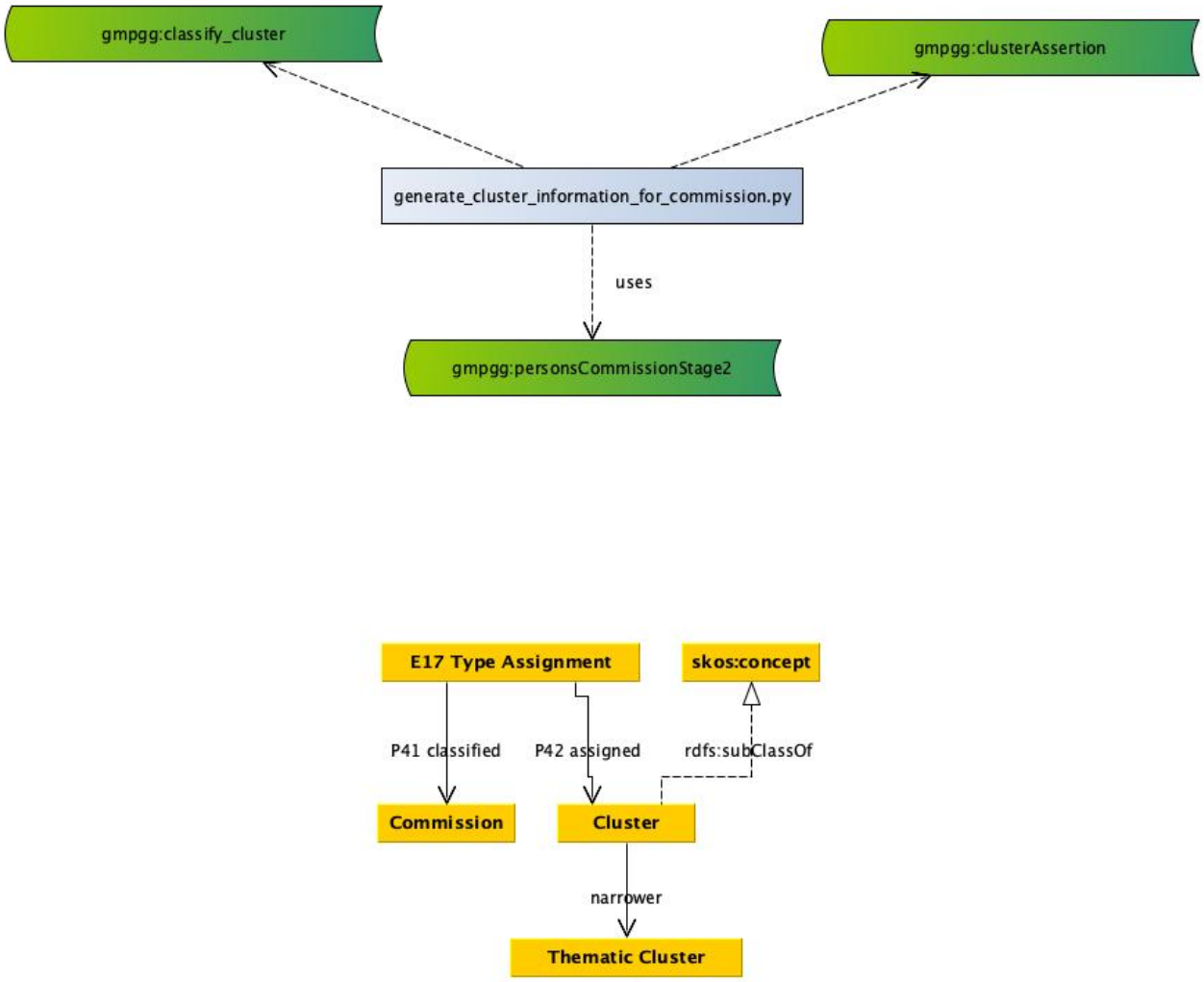


Abbildung 8.27: Zuweisung der thematischen Cluster: unten Struktur der Ontologie, oben Zusammenhang zwischen erzeugendem Skript und den Named-Graphen

[51] Das Vokabular liegt unter:*gmpgg:classify_cluster*, die Klassifikation der Kommissionen in *gmpgg:clusterAssertion*.

8.6.3 Eine erste Bewertung

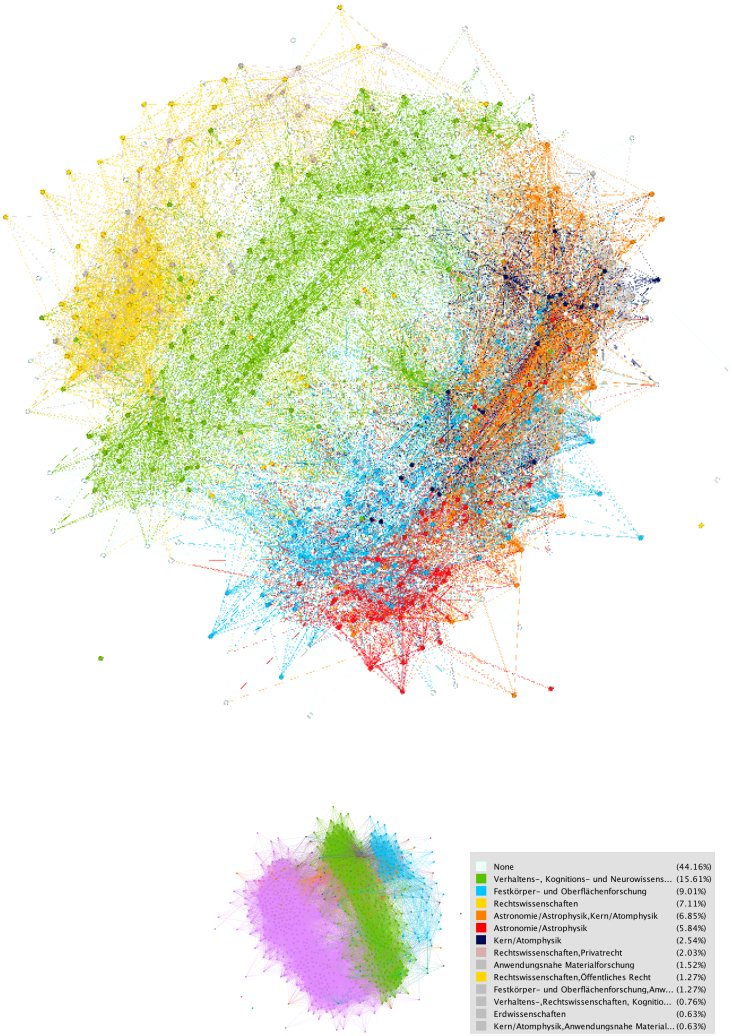


Abbildung 8.28: Kommissionen (ohne AWM) verbunden über gemeinsame Mitglieder, Einfärbung nach Clustern, Darstellung mit Gephi, Fruchtermann-Reingold (Area 10000, Gravity 10.0, Speed 1), unten links zum Vergleich die Einfärbung nach Sektionen

Es ergibt sich auf den ersten Blick (Abbildung 8.28) ein Netzwerk mit sehr unterschiedlichen Regionen. Innerhalb der CPTS finden wir eine Reihe von Clustern, die zum Teil eng miteinander verflochten sind. Kommissionen mit Beteiligung von Personen aus dem Astrobereich (rot) liegen im unteren rechten Bereich des Netzwerks. Darüber finden sich Kommissionen, die sowohl zur Kernphysik als auch zur Atomphysik gehören (orange), sowie darüber ein Bereich der Oberflächenphysik (blau), der je doch stark vor allem in die Kernphysik/Astrophysik hineinragt. Der sehr große Cluster der Verhaltens-, Kognitions- und Neurowissenschaften (grün) durchdringt den gesamten Bereich der BMS und zeigt in der Gesamtansicht keine auffällige Struktur. Die Rechtswissenschaften wiederum grenzen sich als eigene Substruktur innerhalb der GSHS (gelb) deutlich ab. Wir schauen uns als Beispiel das Verhältnis von Astro- und Kernphysik-Cluster etwas genauer an. Abbildung 8.29 zeigt alle Kommission, die

dem Astro-Cluster angehören, Abbildung 8.30 die nur dem Astro-, aber nicht dem Kernphysik-Cluster zuzuordnen sind, und schließlich Abbildung 8.31 Kommissionen, die nur dem Kernphysik-Cluster angehören. Dieser erste Eindruck zeigt, dass der reine Blick auf das Netzwerk, eine Klassifizierung in einen Kernphysik-Cluster ohne die Astrophysik nahelegt, während die andere Einteilung sich im Gesamtbild nicht widerspiegelt.

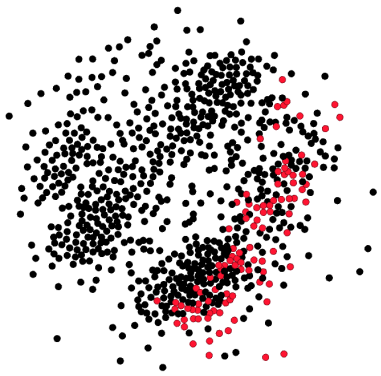


Abbildung 8.29: Lage des Astro-Clusters

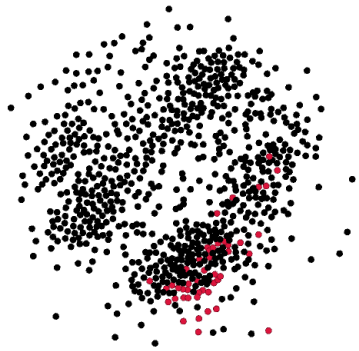


Abbildung 8.30: Kommissionen, die nur dem Astro- aber nicht dem Kernphysik-Cluster angehören.

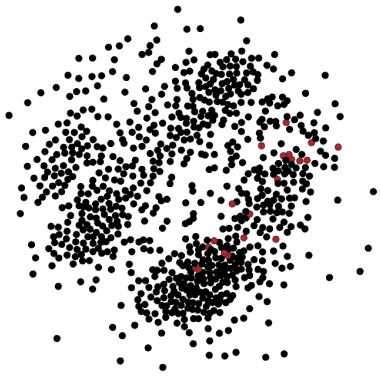


Abbildung 8.31: Kommissionen, die nur dem Kernphysikcluster angehören

8.6.4 Jahresgraphen

Wie beim Personennetzwerk analysieren wir auch hier die Entwicklung des Graphen über die Jahre näher. Erzeugt werden in diesem Fall drei unterschiedliche Jahresgraphen:[52]

- Alle Kommissionen: *doi:21.11103/dataverse.YJL9C0*/com_gr_all_no_simpl.ygz
- Nur Kommissionen, die Cluster verbinden:
  doi:21.11103/dataverse.YJL9C0/com_gr_mul_no_simpl.ygz
- Nur Kommissionen innerhalb von Clustern:
  doi:21.11103/dataverse.YJL9C0/com_gr_no_mul_no_simpl.ygz

Für einen ersten Eindruck schauen wir näher auf die Entwicklung des Verhältnisses der Kanten, die zwischen Kommissionen im gleichen Cluster bzw. zu unterschiedlichen Clustern Verbindungen herstellen (Abb. 8.32).[53]

[52] Notizbuch: *generate_graphs_commissions.ipynb*
[53] *generate_graphs_commissions.ipynb*

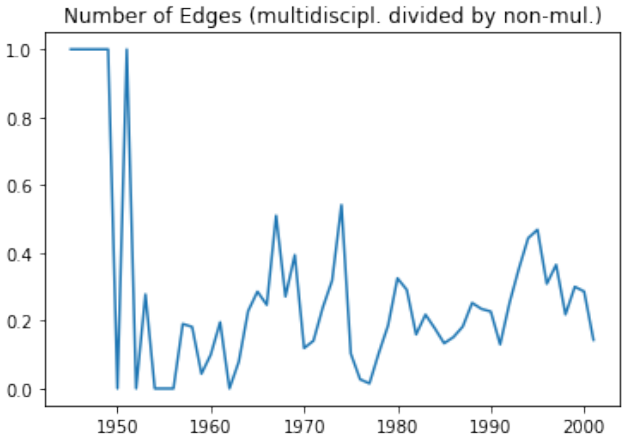


Abbildung 8.32: Verlauf Kantenverhältnis unterschiedliche Cluster / gleich Cluster - mit ex-officio

Vergleichen wir den Verlauf der Charakteristiken des Netzwerks im Verhältnis der CPTS zur BMS[54] (Abb.8.33), so sehen wir ähnliche Periodisierungen, wie wir sie in auch in den Personennetzwerken (Abb. 8.12) beobachten können.[55]

Zur interaktiven Analyse der Graphen steht wieder ein Notebook *Clustering-Kommissionen.ipynb* zur Verfügung. Dieses erlaubt es, Manipulationen an den Graphen vorzunehmen, z. B. die Einschränkung auf bestimmte Knoten. Hiermit lassen sich direkt einzelne Vermutungen über das Netzwerk und Auffälligkeiten in dessen Struktur überprüfen. Eine erste Frage hierbei ist die Interdisziplinarität der Kommissionen. Die Kommissionen wurden den Sektionen zugeordnet, daher lässt sich so ein Hinweis dafür gewinnen, wie stark intersektionelle und damit interdisziplinäre Kooperation in der MPG verankert ist.

[54] *Verlauf-Kommissionen-Gesamt.ipynb*, Daten in *doi:21.11103/dataverse.HVXQTB/*

[55] D.h. eine erste Phase bis etwa 1968, dann eine neue Phase von 1968–1989 und und schließlich eine nach 1989. Auch hier gilt wieder, dass wir die ersten Jahre bis 1955 wegen der geringen Anzahl von Kommissionen als nicht aussagekräftig einschätzen.

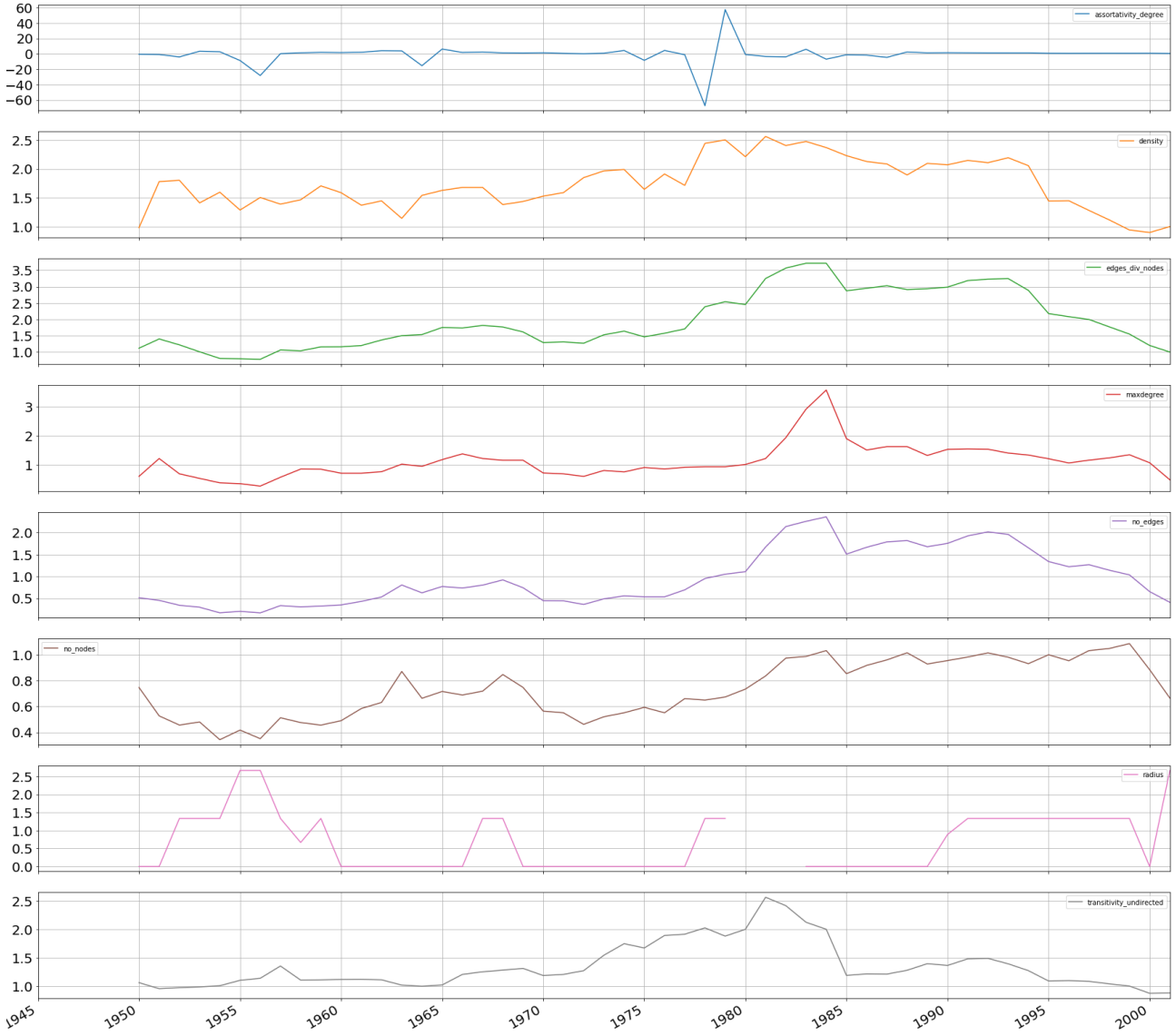


Abbildung 8.33: Entwicklung des Kommissionsnetzwerkes im Verhältnis der Kommissionen

Stellen wir zunächst die Netzwerke aller Kommissionen über die Jahresverläufe in einer Übersicht dar (Abb. 8.35), sehen wir auch hier wieder deutlich eine Umbruchphase analog zu den Beobachtungen im Personengraphen nach 1968, in dem sich die beiden Sektionen deutlich auseinander bewegen. Hierbei fällt zugleich auf, dass die Annahme, dass die Ex-officio-Mitglieder das Bild deutlich verändern, sich ebenfalls bestätigt. Die Abbildung zeigt nur Kommissionen, die mindestens mit einer anderen Kommission aus einer anderen Sektion verbunden sind, und fasst zwei Jahreszeiträume zusammen. Die Umbruchphasen treten hier ohne die Ex-officio-Mitglieder deutlicher hervor.

Die Frage liegt nahe, welche Kommissionen die wenigen Verbindungen in der Zeit nach 1968 darstellen. Der historischen Betrachtung folgend liegt nahe, dass es sich hierbei um die Einrichtung von Instituten und Berufungen in den Überlappungsbereichen zwischen den Sektionen handelt, dies betrifft insbesondere die Biomedizin, Biophysik und medizinische Forschung. Diese offensichtliche Vermutung wird bestätigt, wenn wir die Kommissionen, in denen „biomed“, „biophys“, „medizinische forschung“ vorkommen, mit Hilfe unseres Notizbuches kennzeichnen und in Cytoscape einfärben

(Abb. 8.34). Wir sehen, dass damit fast alle Knoten an den Überlappungsbereichen erfasst werden. Eine weitere Gruppe wird durch Knoten gebildet, die sich mit der Neugründung bzw. Ausgründung eines Institutes zur Linguistik beschäftigen.

Noch deutlicher wird die Bedeutung dieser wenigen Gruppen von Knoten, wenn wir diese aus der Darstellung herausnehmen und wiederum nur die Kommissionen darstellen, die eine Verbindung zwischen den Sektionen darstellen. Es ergibt sich vor allem in den 60er und frühen 80er Jahren (Abb. 8.36) ein eindeutiges Bild: Es existieren fast keine Verbindungen zwischen den Sektionen mehr.

Abbildung 8.34: Jahresentwicklung der Kommissionen (ohne ex-officio), nur Kommissionen, die zu Kommissionen aus anderen Sektionen verbunden sind.

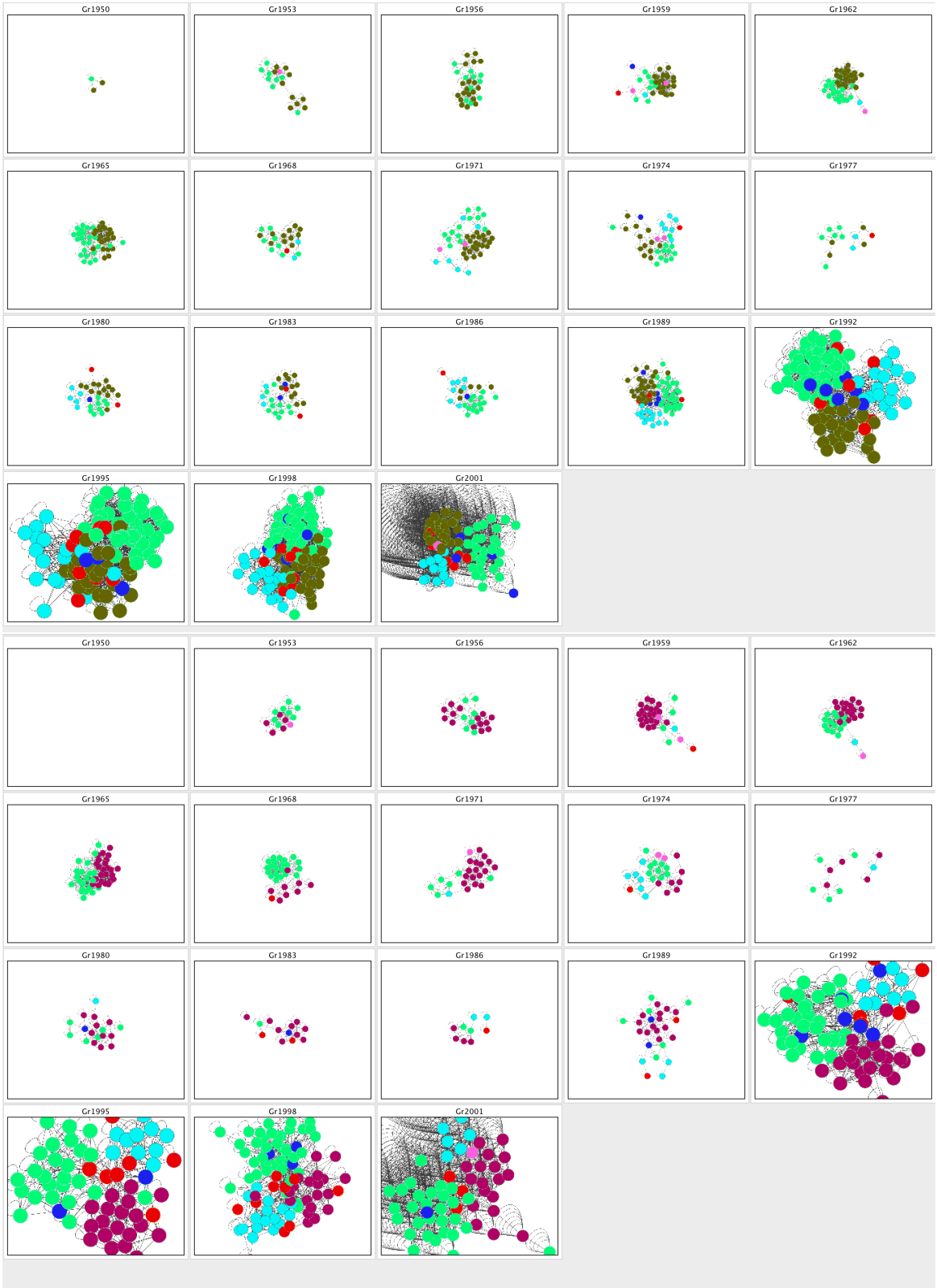


Abbildung 8.35: Jahresentwicklung der Kommissionen (ohne ex-officio unten), nur Kommissionen, die mit Kommissionen aus anderen Sektionen verbunden sind.

Abbildung 8.36: Jahresentwicklung der Kommissionen (ohne ex-officio) - ohne „biomed“„biophys“„medizinische forschung“ und Linguistik, nur Kommissionen, die zu Kommissionen aus anderen Sektionen verbunden sind.

8.7 Cluster und Cluster

Die nächste Frage ist die nach der Rolle der im Forschungsprojekt als Arbeitshypothesen definierten inhaltlichen Cluster. Lassen sich diese Cluster in den Graphen wiederfinden? In Abschnitt 8.6.3 haben wir bereits eine erste Bewertung der Rolle der Cluster mittels der graphischen Darstellung vorgenommen. Wir wollen dies nun etwas vertiefen. Dazu untersuchen wir die Dynamik der Clusterentwicklung (Abb. 8.37) und suchen nach Korrelationen zwischen Veränderungen der Clusterstruktur und Charakteristiken des Netzwerks in der zeitlichen Entwicklung. Auch diese Analysen sind in einem Notizbuch[56] zusammengefasst.

In erster Linie vergleichen wir das Ergebnis von netzwerktheoretischen Clusteranalysen (SNA-Cluster) mit der inhaltlich-strukturell angenommenen Clusterstruktur. Dazu betrachten wir die Korrelation von inhaltlichen Clustern und SNA-Clustern. Abb. 8.38 zeigt den Prozentsatz der Kommissionen eines inhaltlichen Clusters, die in einem SNA-Cluster liegen. In den Jahren, in denen wenige Punkte dicht übereinander liegen, bilden die inhaltlichen Cluster auch SNA-Cluster.

Ein weiterer Hinweis auf die Eigenschaft der inhaltlichen-strukturellen Cluster, auch im Rahmen der SNA als distinktive Eigenschaft zu dienen, ist die Zahl, wie oft mehr als ein bestimmter Anteil der inhaltlichen Cluster in einem einzelnen SNA-Cluster liegen. Hier sehen wir in der Gesamtansicht, dass der Kernphysik-Cluster und der Privatrechts-Cluster die am stärksten ausgeprägten Cluster sind: Hier liegen mehr als die Hälfte der Zeit 90% der Kommissionen in einem Cluster (Abb. 8.39, rechts). Für die überwiegende Zahl aller Cluster ist das Bild nicht so deutlich, jedoch liegen auch hier in der Regel

[56] *Clustering-Kommissionen-Indikatoren- Qualität_der_Cluster.ipynb*

mehr als die Hälfte aller Kommissionen in einem SNA-Cluster über den Zeitverlauf (Abb. 8.39, links). Ein differenzierteres Bild bekommen wir in der zeitlichen Auflösung (Abbi. 8.40). Zum Vergleich sind auch die Ergebnisse bezüglich anderer Clusteringverfahren (Abbildungen 8.42; 8.41) angegeben.

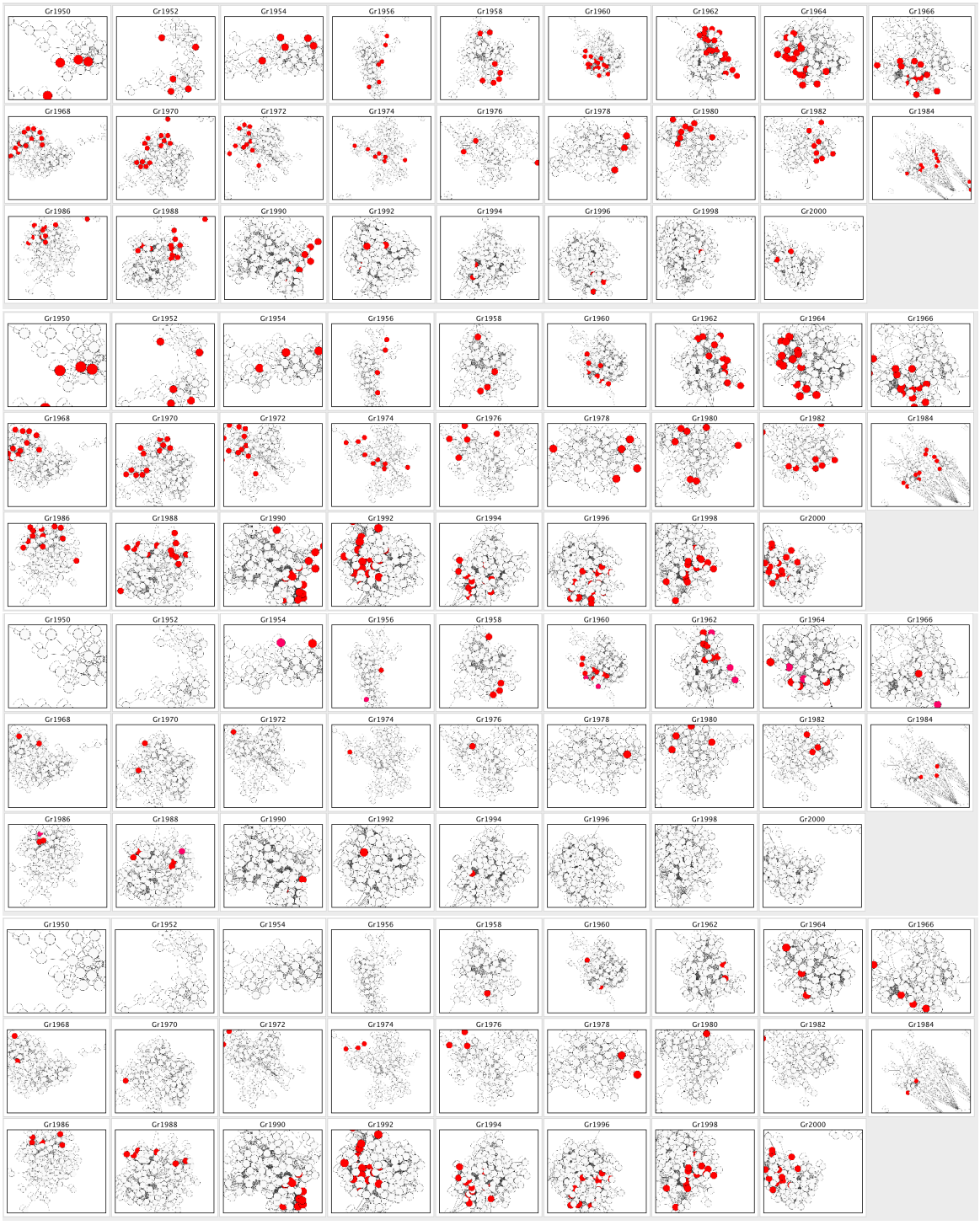


Abbildung 8.37: Jahresentwicklung der Kommissionen, der Reihe nach Kernphysik, Astrophysik, Kernphysik ohne Astrophysik, Astrophysik ohne gemeinsame Kommissionen mit Kernphysik für jeweils 5 Jahre zusammengefasst.

Reduzieren wir die Graphen nur auf persönlich benannte Mitglieder, d. h. ohne die Ex-officio-Mitglieder, so nimmt der prozentuale Anteil für alle Cluster ab – auffällig ist jedoch der relativ starke

Abfall des Clusters aus der Kernphysik (Abbildung 8.43 und 8.44). In der zeitlichen Entwicklung ergibt sich jedoch der gleiche Trend wie zuvor.

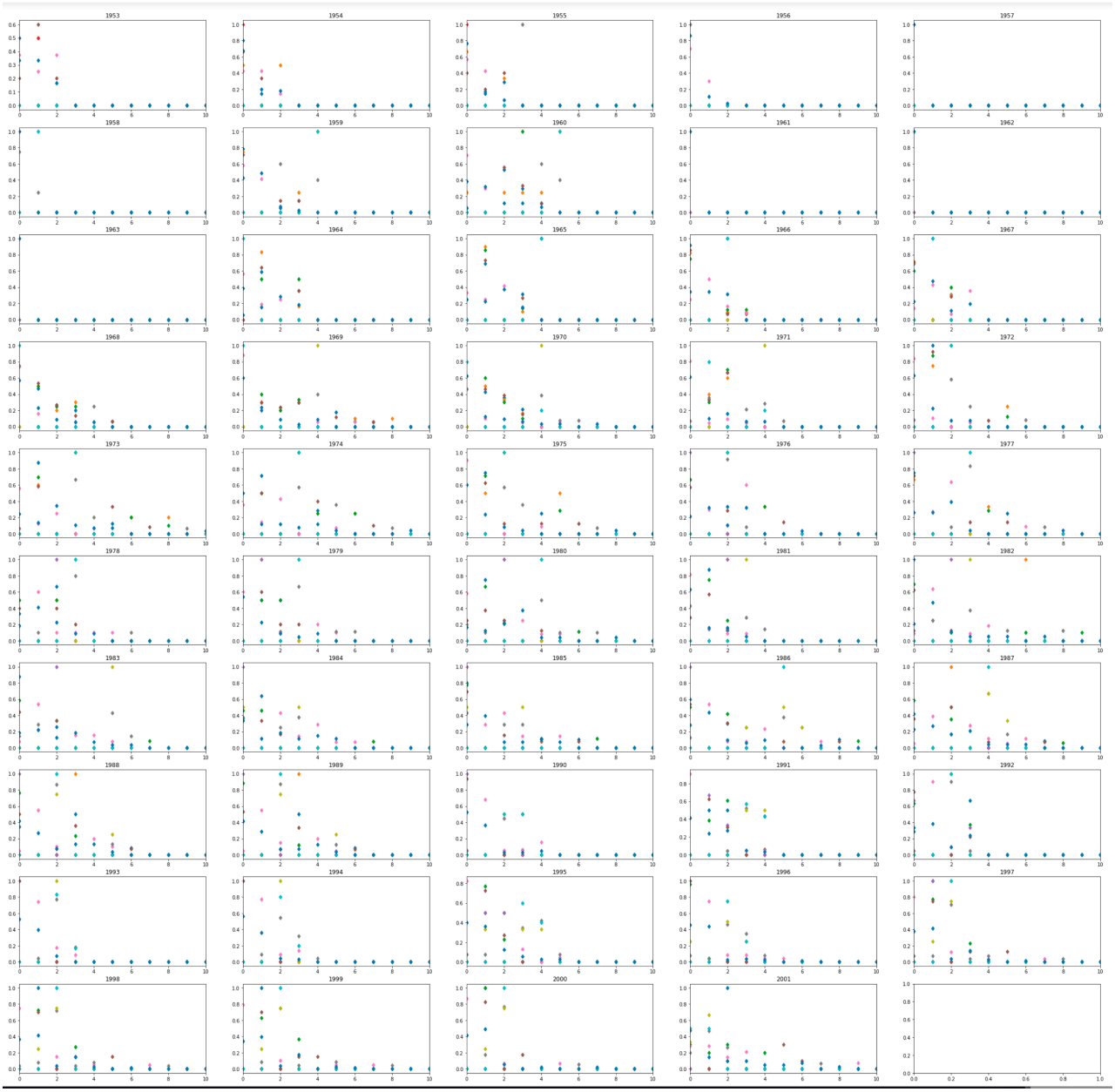


Abbildung 8.38: Verteilung der inhaltlichen Cluster über automatische generierten (infomap).

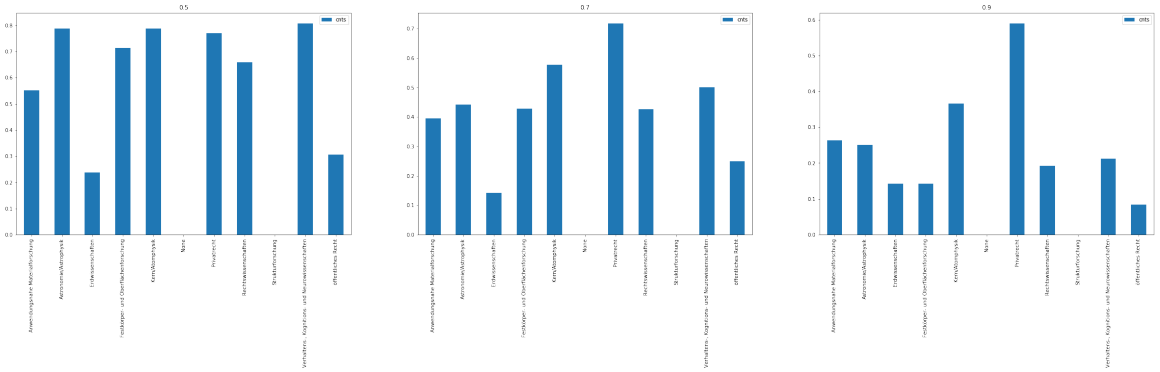


Abbildung 8.39: Prozentualer Anteil, wie oft ein Cluster mit mehr als einem Threshold von 0.5,0.7,0.9 über die Jahre in einem SNA-Cluster liegt.

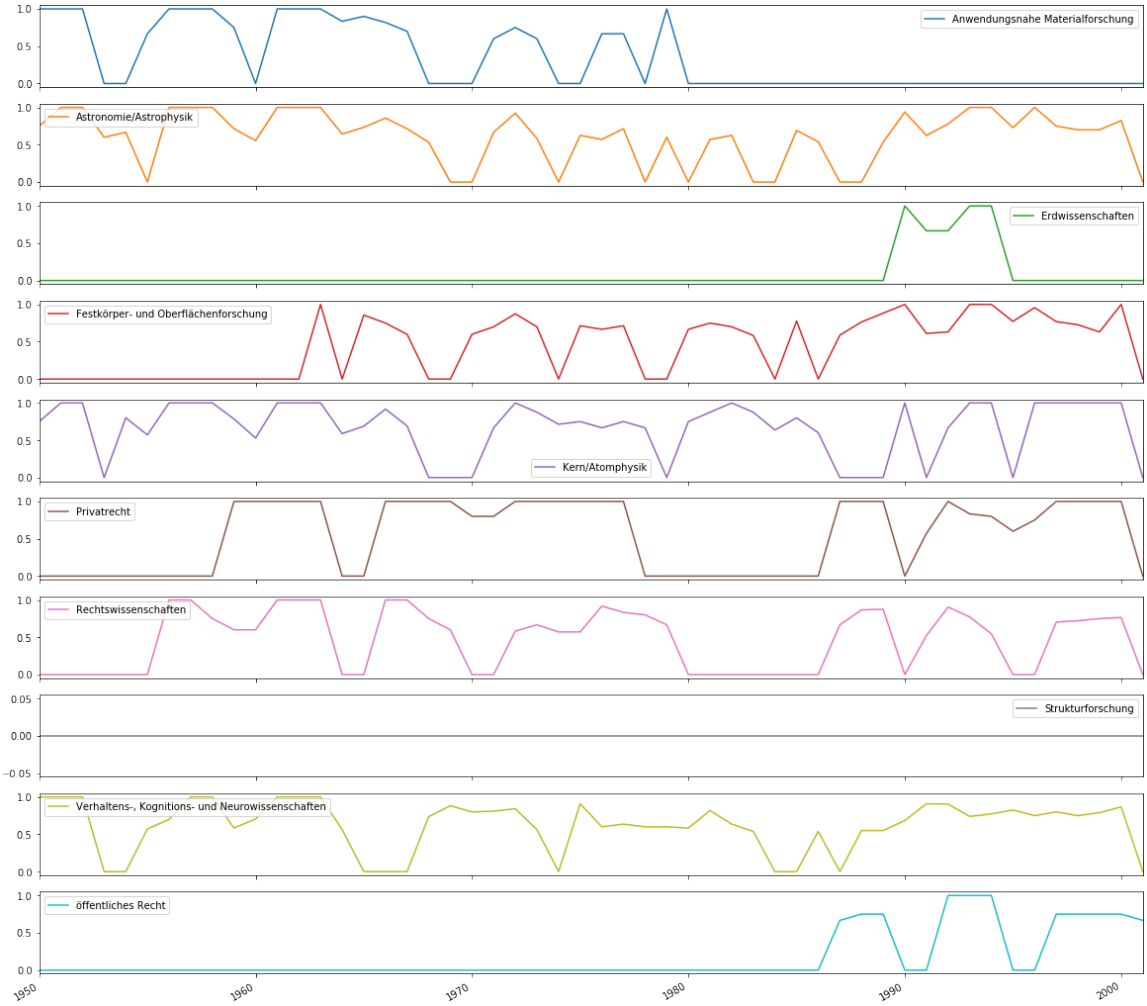


Abbildung 8.40: Zeitliche Auflösung - Anteil der Cluster mit mehr als 50% in einem SNA-Cluster mit Infomap

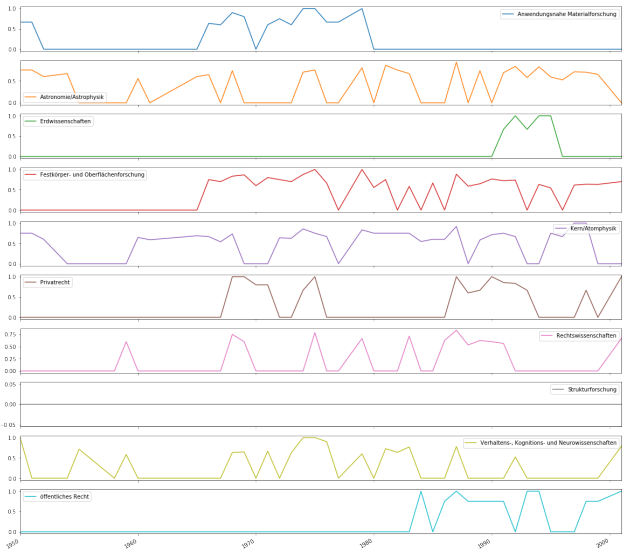


Abbildung 8.41: Zeitliche Auflösung - Anteil der Cluster mit mehr als 50% in einem SNA-Cluster nach Betweeness

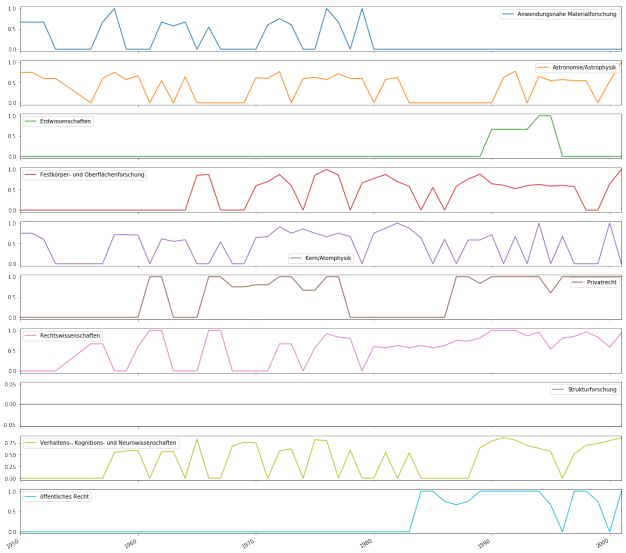


Abbildung 8.42: Zeitliche Auflösung - Anteil der Cluster mit mehr als 50% in einem SNA-Cluster mit Leading Eigenvektor.

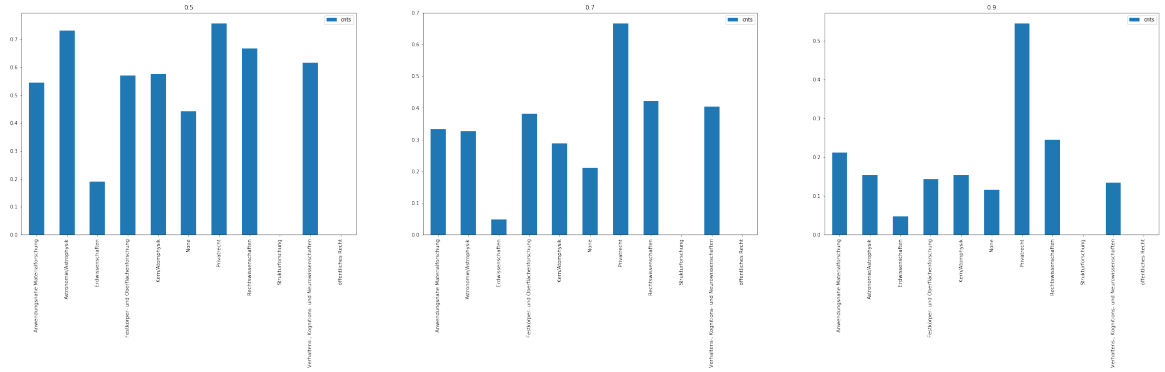


Abbildung 8.43: Anzahl, wie oft ein Cluster mit mehr als einem Threshold von 0.5,0.7,0.9 über die Jahre in einem SNA-Cluster liegt - ohne ex-officio - infomap

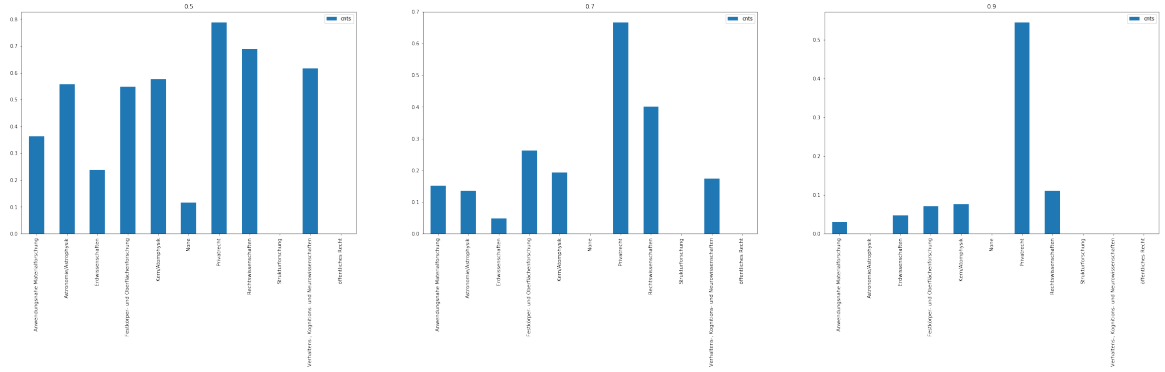


Abbildung 8.44: Anzahl, wie oft ein Cluster mit mehr als einem Threshold von 0.5,0.7,0,0 über die Jahre in einem SNA-Cluster liegt - ohne ex-officio - leading eigenvectors

Vergleichen wir die zeitliche Entwicklung der Netzwerke mit allen Mitgliedern (Abb. 8.41) mit den

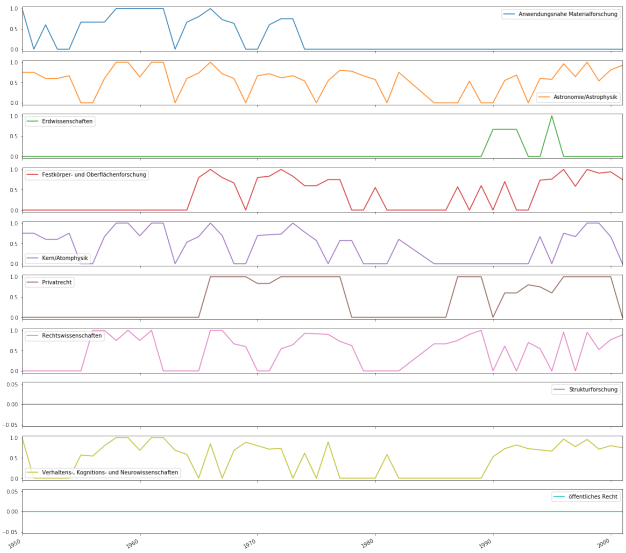


Abbildung 8.45: Zeitliche Auflösung - Anteil der Cluster mit mehr als 50% in einem SNA-Cluster nach Informap ohne Ex-Officio

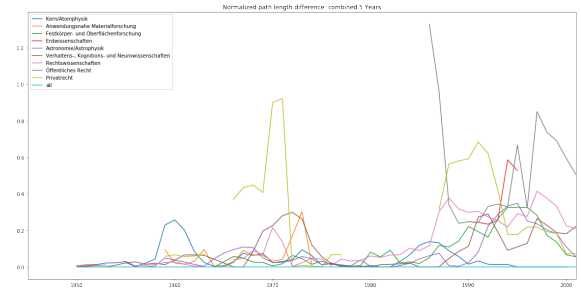


Abbildung 8.46: Entwicklung der Pfadlängen innerhalb der Cluster (5 Jahre)

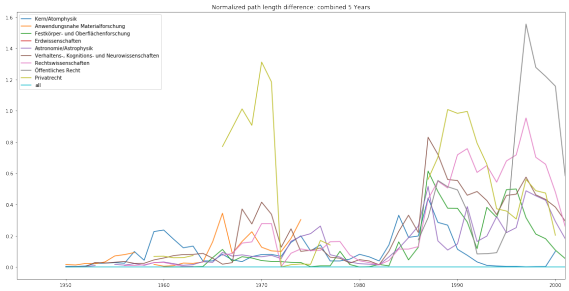


Abbildung 8.47: Entwicklung der Pfadlängen innerhalb der Cluster (5 Jahre) - ohne ex-officio

Netzwerken ohne die Ex-officio-Mitglieder (Abb. 8.45), so wird der Grund für die Abweichungen deutlich. Einbrüche sind fast immer eine Folge davon, dass ein Mitglied des Clusters zum Sektionsvorsitzenden gewählt wird, so z.B. im Jahr 1988 der Direktor am MPI für Kernphysik, Hans-Arwed Weidenmüller. Diese Beobachtungen stehen in Einklang mit den Beobachtungen in 8.5.5. Sektionsvorsitzende spielen in der Regel vor ihrer Wahl bereits eine Rolle in den Sektionen, so dass ihre Nichtbeachtung in den Netzwerken eine deutliche Veränderung des Netzes nach sich zieht. Das Gewichtungsproblem der institutionellen Mitglieder in Kommissionen wird dadurch erneut deutlich.

Wir betrachten zusätzlich einige weitere Indikatoren: die Entwicklung der durchschnittlichen Weglänge, die Kommissionen desselben Clusters verbindet, und die Modularität. Die Graphen, die jeweils fünf Jahre[57] umfassen, zeigen vier unterschiedliche Phasen: vor 1959, 1959–1969, 1969–1980 sowie nach 1980. Diese sind jeweils auf der einen Seite durch einen Wechsel der Stärke der Modularität gekennzeichnet (Abb.8.48 und 8.49) und auf der anderen Seite durch einen Wechsel des relativen Verhältnisses der durchschnittlichen Pfadlängen in den Teilnetzwerken. Aufgrund der impliziten Ab-

[57] Zum Vergleich auch die Graphen, die 2 Jahre zusammenfassen Abb. 8.52 und 8.53, diese zeigen eine ähnliche Entwicklung.

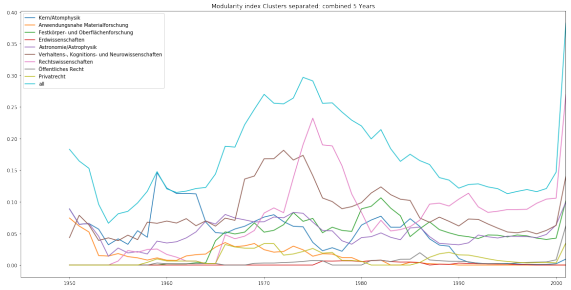


Abbildung 8.48: Modularitätsmaß für die Cluster (5 Jahre)

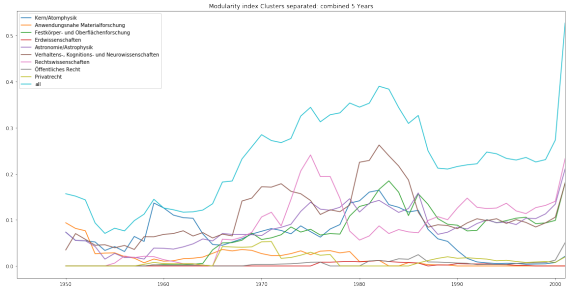


Abbildung 8.49: Modularitätsmaß für die Cluster (5 Jahre) - ohne ex-officio

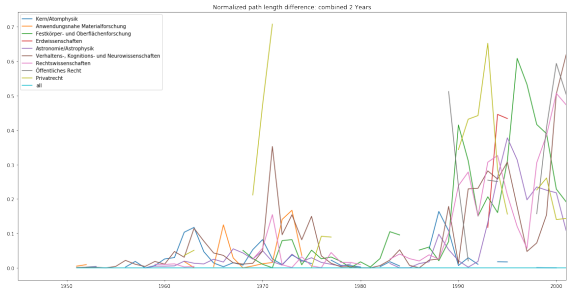


Abbildung 8.50: Entwicklung der Pfadlängen innerhalb der Cluster (2 Jahre)

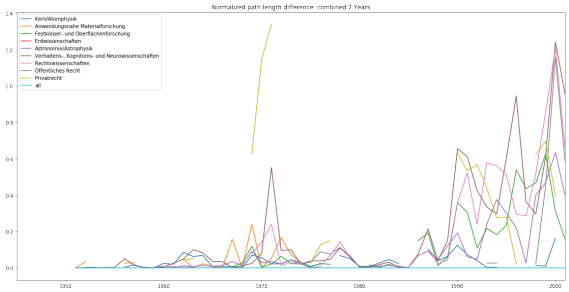


Abbildung 8.51: Entwicklung der Pfadlängen innerhalb der Cluster (2 Jahre) - ohne ex-officio

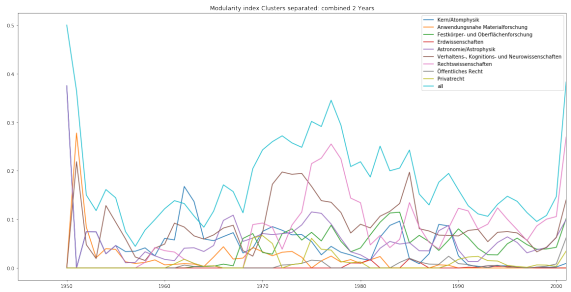


Abbildung 8.52: Modularitätsmaß für die Cluster (2 Jahre)

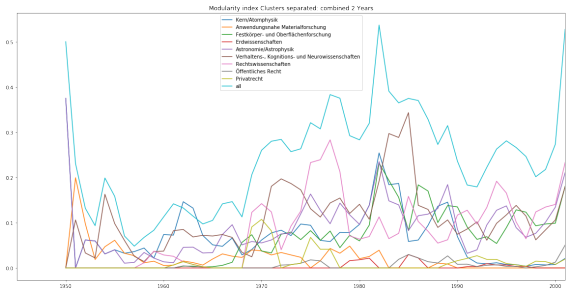


Abbildung 8.53: Modularitätsmaß für die Cluster (2 Jahre) - no ex officio

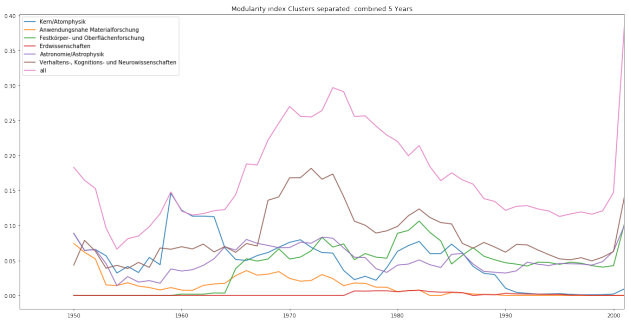


Abbildung 8.54: Modularitätsmaß für die Cluster (5 Jahre) - Ohne Rechtswissenschaften

hängigkeit der Modularität von den Pfadlängen ist zu erwarten, dass im Zeitraum geringer Modularität die Pfadlängen innerhalb der einzelnen Cluster sich in der Nähe des Wertes des Gesamtnetzwerkes aufhalten (Abb.8.50, 8.51, 8.46, 8.47). Interessant ist jedoch, dass sich die einzelnen Cluster in den anderen Phasen deutlich unterschiedlich entwickeln. Dies gilt vor allem für die Cluster, die mehrheitlich in der CPTS liegen, während der Neuro-Cluster im wesentlichen der Gesamtentwicklung folgt. (8.54).

Um eine Überbewertung der Abweichung von Jahren mit nur einer geringen Anzahl von Kommissionen zu vermeiden, gewichten wir die Abweichungen der mittleren Pfadlängen der einzelnen Cluster mit dem Quadrat der relativen Anzahl der Kommissionen.

$$\mathfrak{n} = \mathfrak{y}^2_{\mathsf{df\_diff}} \left(\frac{\mathfrak{y}_{\mathsf{num\_of\_com}}}{\mathfrak{y}_{\mathsf{num\_of\_com\_max}}} \right)^2$$

Der Verlauf insbesondere der Kurven für den Cluster „Festköper- und Oberflächenphysik“ untermauert die Vermutung der unterschiedlichen Phasen, wie sie von Thomas Steinhauser, einem der Mitarbeiter des Projektes, aufgestellt wurde: Eine relativ lange Findungsphase des Clusters, beginnend mit der Gründung des MPI für Festkörperforschung 1969 bis in die späten 1980er Jahre, danach eine erste Ausprägung des Clusters mit der Gründung des MPI für Polymerforschung 1983 und dann schließlich eine starke Ausprägung des Clusters im Zuge der Neugründung von Instituten nach der deutschen Wiedervereinigung.

8.8 Koinfluenz

Mit den Koinfluenznetzwerken verfolgen wir, wie schon einleitend beschrieben[58], einen Ansatz analog zu den Kozitationsnetzwerken, die wir bei der Fallstudie zur ART untersucht hatten.[59] Diese beruhte auf der Annahme, dass Kozitationen Einsichten darüber ermöglichen, wie durch die Autoren Forschungsstände so aufgenommen und neu kombiniert werden, dass dadurch neue Forschungsbereiche aufgetan werden. Unsere Hypothese ist hierbei, dass sich die Inhalte der Kommissionen in dieser Interpretation ähnlich verstehen lassen. Kommissionen bringen durch ihre Akteure Felder zusammen, die vorher nicht zusammengebracht wurden, und führen schließlich zu Instituten, die neue Themenbereiche verfolgen. Im Sinne des zu Beginn eingeführten Ansatzes, in einer Theorie der Wissensorganisation unterschiedliche Ebenen des Wissensnetzwerkes zusammenzuführen, stellen die Kommissionen eine erste mögliche Abstraktion von der Akteursebene zu einer semantischen Ebene dar. Die Kommissionen selbst als formalisierte Systeme des Informationsaustausches wiederum liegen analog zu den veröffentlichten Artikeln auf der Ebene des *semiotischen Netzwerkes*.

Die Frage, der wir hier nachgehen wollen, ist, ob wir mittels eines Netzwerkes, das Kommissionen nicht über gemeinsame Mitglieder, sondern darüber zusammenbringt, dass Mitglieder der Kommission später gemeinsam in einer Kommission saßen, eine ähnliche Entwicklung von Trends erkennen können. Im Gegensatz zu den Netzwerken bisher in diesem Kapitel ist das Ergebnis nun ein gerichtetes Netzwerk. Zusätzlich schließen wir alle Verbindungen zwischen Kommissionen aus, die durch glei-

[58] Siehe Abschnitt 8.6.

[59] Siehe Abschnitt 7.7.

che Mitglieder bereits verbunden sind. Wir betrachten außerdem nur die Mitglieder von Kommissionen, die nicht Ex-officio-Mitglieder sind. Es ergibt sich damit eine der Kozitationsanalyse vergleichbare Situation. Wir nehmen an, dass neue Themenfelder dadurch erschlossen werden, dass eine Kommission Vertreter von Feldern zusammenbringt, die vorher noch nicht kooperiert haben. Auch hier gilt natürlich, dass wir uns des eingeengten Blicks, den die Kommissionen als Quelle erlauben, bewusst sind. Die Aussagekraft der Ergebnisse ist vor diesem Hintergrund sicher eingeschränkt. Ohne einen tieferen Einblick sowohl in die Inhalte der Kommissionen und der Forschungsfelder der Personen, die in den entsprechenden Kommissionen tätig waren, als auch in die sozialen Netzwerke hinter den Kommissionsnetzwerken ist eine belastbare Aussage über die Formierung von Themenfeldern durch die Arbeit der Kommissionen nicht möglich. Jedoch helfen die Aussagen aus den Netzwerkanalysen bei der Identifikation von Bereichen für zukünftige historische Tiefenbohrungen.

Auch hier vergleichen wir die unterschiedlichen Netzwerke wie oben mit verschiedenen Einschränkungen. Wir konstruieren das Netzwerk ausgehend von dem bipartiten Graphen der Kommissionen und Personen.[60] Der Kern zur Erzeugung des Koinfluenznetzwerkes ist im Wesentlichen die Berechnung einer Abhängigkeitsmatrix für jedes Jahr. Dabei iterieren wir über alle Jahre und bestimmen jeweils, ob diese Kommission in der Zukunft mit einer anderen über deren Mitglieder verbunden werden wird.[61]

Vergleichen wir die Abbildungen 8.55 und 8.56, so ist vor allem die deutlich veränderte Rolle der GSHS auffällig. Diese ist jetzt deutlich stärker in die anderen Sektionen integriert – stärker in die BMS als in die CPTS. Außerdem existiert nun ein sehr dichter Überlappungsbereich zwischen GSHS und BMS. Auf den ersten Blick sind auch die Cluster weniger deutlich ausgeprägt.

Wir vergleichen im Folgenden Betweenness-, Closeness- und Degree-Zentralitäten für die beiden Netzwerke. Schauen wir hier auf die Mittelwerte dieser Zentralitäten für die unterschiedlichen Typen von Kommissionen, d.h. Gründungs-, Berufungs- und Zukunftskommissionen,[62] so erhalten wir die in den Abbildungen 8.57 und 8.58 ersichtlichen Veränderungen. Im wesentlichen steigt die Bedeutung der Zukunftskommissionen an. Auch wenn wir auf die Rolle einzelner Kommissionen schauen, steigt die Bedeutung von Kommissionen, die sich mit neuen Feldern beschäftigen, leicht an. Auffällig ist diese besonders bei Closeness- und Betweenness-Zentralität: Hier liegen jetzt Zukunftskommissionen auf den ersten Plätzen und die Kommission zur Gründung eines Musikinstitutes führt die Rangliste in Bezug auf Betweenness-Zentralität. Im Vergleich dazu liegt diese Kommission im Kooperationsnetzwerk lediglich auf Rang 473.

8.9 Zusammenfassung und Ausblick

Die ersten Analysen geben uns einen Eindruck über die innere Struktur des Kommissions- und Personennetzwerkes. Einzelne zentrale Personen lassen sich identifizieren, insbesondere diejenigen, die zwischen einzelnen Themenfeldern stehen. Wir sehen auch, wie erwartet, dass die einzelnen Sek-

[60] *21.11103/dataverse.GIUCSK*/commissionsAndPersons.graphml

[61] *co-influence-commissions.ipynb*

[62] *Verlauf-Kommissionen-Koinfliuence.ipynb*,*Verlauf-Kommissionen.ipynb*

	betw	betw_name	close	close_name	degree	degree_name
0	15097.8	Ständige Wissenschaftliche Kommission zur Über...	0.221878	Ständige Wissenschaftliche Kommission zur Über...	677	Kommission „Forschungsperspektiven der BMS“
1	7004.51	Kommission „Wissenschaftsgeschichte“/ Berufung...	0.212818	Präsidentenkommission „Großprojekte in der MPG“	645	Zweiter Intersektioneller Arbeitskreis Systeme...
2	6360.53	Präsidentenkommission „Großprojekte in der MPG“	0.212703	Zweiter Intersektioneller Arbeitskreis Systeme...	643	Ständige Wissenschaftliche Kommission zur Über...
3	5917.16	Gründung eines MPI für theoretische Biologie	0.212244	Stammkommission MPI für Metallforschung / Beru...	638	Zukunft des MPI für Strömungsforschung / Nachf...
4	5534.12	Zweiter Intersektioneller Arbeitskreis Systeme...	0.212072	Erweiterung des Direktoriums des MPI für Astro...	598	Nachfolge Paneth / Berufungen an das MPI für C...
5	5042.91	Bericht „Wissenschaftsrat“ / Empfehlungen des ...	0.211673	Vorschlag zur Gründung eines MPI für Physik ko...	596	Stammkommission MPI für Metallforschung / Beru...
6	4840.99	Erweiterung des Direktoriums des MPI für Astro...	0.211616	Gründung eines MPI für theoretische Biologie	578	Strategie-Kreis der CPT-Sektion
7	4759.98	Zukunft des MPI für physikalische Chemie / Zen...	0.211616	Berufung v. Prof. Dr. K. BInder / eines WM und...	532	Präsidentenkommission „Großprojekte in der MPG“
8	4662.17	Ernennung von Herbert Jäckle /Jaeckle zum WM d...	0.211162	Ergänzung der wiss. Leitung des MPI für Plasma...	507	Innovation und Innovationsfähigkeit innerhalb ...
9	4408.3	Ernennung von Dr. Jovin zum WM des MPI für bio...	0.21071	Gründung MPI für Informatik / Wiedereinsetzung...	505	Ergänzung der wiss. Leitung des MPI für Plasma...
10	4301.85	Gründung MPI für Informatik / Wiedereinsetzung...	0.210597	Zukunft des MPI für physikalische Chemie / Zen...	477	Verselbstständigung der Teilinstitute des MPI ...

Tabelle 8.4: Verteilung auf Typen, Mittelwerte und absolute Werte (abw) und Standardabweichung (abw)

	betw	betw_name	close	close_name	degree	degree_name
0	3634.57	Intersektionelle Kommission „Musik-Institut“ /...	0.0202693	Zukunft des MPI für Verhaltensphysiologie, ins...	544	Zukunft des MPI für Verhaltensphysiologie, ins...
1	2857.55	Ergänzung der wiss. Leitung des MPI für Plasma...	0.0202528	Zukunft des MPI für Biochemie (Emeritierung Ly...	508	Zukunft des MPI für Biochemie (Emeritierung Ly...
2	2747.09	Zukunft des MPI für Verhaltensphysiologie, ins...	0.0202491	Ergänzung d. wissensch. Leitung / Ernennung vo...	500	Ergänzung d. wissensch. Leitung / Ernennung vo...
3	2503.21	Berufung von Prof. Hans Kuhn in die MPG	0.0202486	Berufung Dr. K. G. Götz zum Direktor am MPI fü...	499	Berufung Dr. K. G. Götz zum Direktor am MPI fü...
4	2079.66	Nachfolge Kramer / Ernennung von Mittelstaedt ...	0.0202477	Berufung Hartmut Michel zum WM und Mitglied de...	497	Berufung Hartmut Michel zum WM und Mitglied de...
5	2076.28	Ernennung von Dr. Jovin zum WM des MPI für bio...	0.0202468	Berufung Prof. Jan Klein zum WM und Direktor e...	495	Berufung Prof. Jan Klein zum WM und Direktor e...
6	1968.55	Berufung Dr. K. G. Götz zum Direktor am MPI fü...	0.0202463	Zukunft des MPI für Himforschung (Emeritierun...	494	Zukunft des MPI für Himforschung (Emeritierun...
7	1942.53	Ständige Wissenschaftliche Kommission zur Über...	0.0202459	Bibliotheks-Kommission (Informations- und Lite...	493	Bibliotheks-Kommission (Informations- und Lite...
8	1928.45	Erweiterung des Direktoriums des MPI für Astro...	0.0202449	Kommission „Linguistik“ /Teilnachfolge der Ges...	491	Kommission „Linguistik“ /Teilnachfolge der Ges...
9	1858.6	Ernennung von Prof. R. v. Baumgarten zum WM am...	0.0202449	Berufung Prof. Dr. Holger Schmid-Schönborn zum...	491	Berufung Prof. Dr. Holger Schmid-Schönborn zum...
10	1788.57	Kommission „Linguistik“ /Teilnachfolge der Ges...	0.020244	Intersektionelle Kommission „Musik-Institut“ /...	489	Intersektionelle Kommission „Musik-Institut“ /...

Tabelle 8.5: Verteilung auf Typen, Mittelwerte und absolute Werte(co-influence)

tionen in der Regel deutlich voneinander abgegrenzt sind und sich nur dort überlappen, wo bewusst interdisziplinäre bzw. sektionsübergreifenden Themen behandelt wurden. Die vergleichenden Untersuchungen der beiden Netzwerkstrukturen, die wir zwischen Kommissionen bilden können, geben Hinweise auf Kanten, die in einer nun noch durchzuführenden komplexeren Analyse mit Informationsflüssen unterschiedlicher Art verbunden werden können. Mit den in der Fallstudie geschilderten Methoden können wir eine Reihe zentraler Fragen noch nicht beantworten. So haben wir nur ein sehr ungenaues Bild über die in den Kommissionen tatsächlich verhandelten Punkte. Hier benötigen wir einerseits eine Analyse der vollständigen Protokolle der Kommissionen. Dieses kann teilweise durch Textmining geschehen, wird aber eine Detailstudie nicht ersetzen. Hier werden wir durch einzelne, sich aus den Ergebnissen der geschilderten Studie ergebene Schwerpunktsetzungen gezielte Fallstudien vornehmen. Es zeigt sich, dass die Cluster sich in den Netzwerken durchaus als ordnende Strukturen zeigen – wenn auch in sehr unterschiedlicher Stärke. Auch die historischen Periodisierungen sind in den einzelnen Verläufen der Entwicklung nachvollziehbar.

Der in dieser Fallstudie geschilderte Ansatz ist mit der Konzentration auf die Kommissionen stark eingeschränkt und diente im Rahmen dieser Arbeit hauptsächlich dazu, Methoden auszutesten. Am Beispiel der Mandatsträger in der MPG sehen wir, dass wir genauere Modelle dafür entwickeln müssen, wie diese auf Entscheidungsstrukturen Einfluss nehmen. So sehen wir in den Netzwerken keine eindeutigen Zentralisierungstendenzen. Dies ist jedoch auch zunächst nicht überraschend, da wir über die Netzwerke zwar ableiten können, wer an Entscheidungen beteiligt ist, jedoch letztendlich nicht sehen können, auf welcher Grundlage dann Entscheidungen getroffen wurden. Wir können auch nicht ableiten, wie Anstöße zu Kommissionen gegeben wurden. Dazu bedarf es noch wesentlich genauerer Studien darüber, wie Entscheidungsvorlagen in die einzelnen Kommissionen und Gremien Eingang gefunden haben. Um dies aus den Quellen ablesen zu können, bedarf es noch genauerer Hypothesen über die Handlungsabläufe, die den Gremiensitzungen vorangegangen sind. Diese Hypothesen können dann ein Ausgangspunkt für das Text- und Datamining sein, in dem etwa näher untersucht wird, wer zu welchen Zeitpunkten und in welcher Form Punkte in die Gremien eingebracht hat. So ist die Frage nach Zentralisierungsvorgängen mit unseren Analysen bisher nicht zu beantworten, ohne mit expliziteren Modellen zu bestimmen, was Zentralisierung im Kontext der MPG bedeutet.

Unser allgemeiner theoretischer Ansatz einer Multileveltheorie der Wissensentwicklung geht von einer starken Interpendenz der unterschiedlichen Handlungs- und Wissensebenen aus. Externe Faktoren der Wissensentwicklung, ihr gesellschaftlicher und politischer Rahmen und die Entwicklung von Forschungsfeldern haben wir bisher in den Netzwerkanalysen noch nicht berücksichtigt. Hier steht eine Verbindung der ersten Fallstudie mit dieser noch an, in der wir die Entwicklung ausgesuchter Forschungsfelder mit der internen Entwicklung der MPG vergleichen. Das Wechselspiel unserer Netzwerke mit anderen Netzwerkstrukturen wie dem komplexen *semiotischen Netzwerk* von experimentellen Methoden sowie die Institutionalisierung von Forschung in anderen Strukturen müssen noch weiter untersucht werden. Konkretes Beispiel dafür könnte die „Computerisierung“ der Forschung sein, d. h. der Einzug von Auswertungs- und Messmethoden, die dank der rasant steigenden Leistungsfähigkeit bei gleichzeitig sinkenden Investitionskosten mit Hilfe von Computern durchgeführt werden können.

Auf der rein methodischen Ebene steht noch die Auswertung der Modellierung der Netzwerke

mittels ERGM und RSIENA. Erste Versuche dazu haben wir mittlerweile unternommen, jedoch sind die Ergebnisse noch zu unzuverlässig, um sie hier vorzustellen. Auch dies ist eng mit dem Problem verbunden, dass wir noch keine überzeugenden Modelle dafür entwickeln konnten, welche Bindungen wir tatsächlich erwarten. Die schrittweise Beantwortung der genannten Fragen nach den Handlungsstrukturen wird dabei helfen, die Modelle zu verbessern und umgekehrt zu Grunde liegende Hypothesen zu überdenken.

Schließlich haben wir begonnen, Clusteranalysen auf der Grundlage von Multilevel-Netzwerken vor allem mit Infomap durchzuführen. Auch dort sind wir noch nicht zu interpretierbaren Ergebnissen gekommen; die genaue Schilderung der Schritte, die wir dort bisher unternommen haben, würde zugleich den Rahmen dieser Arbeit sprengen.

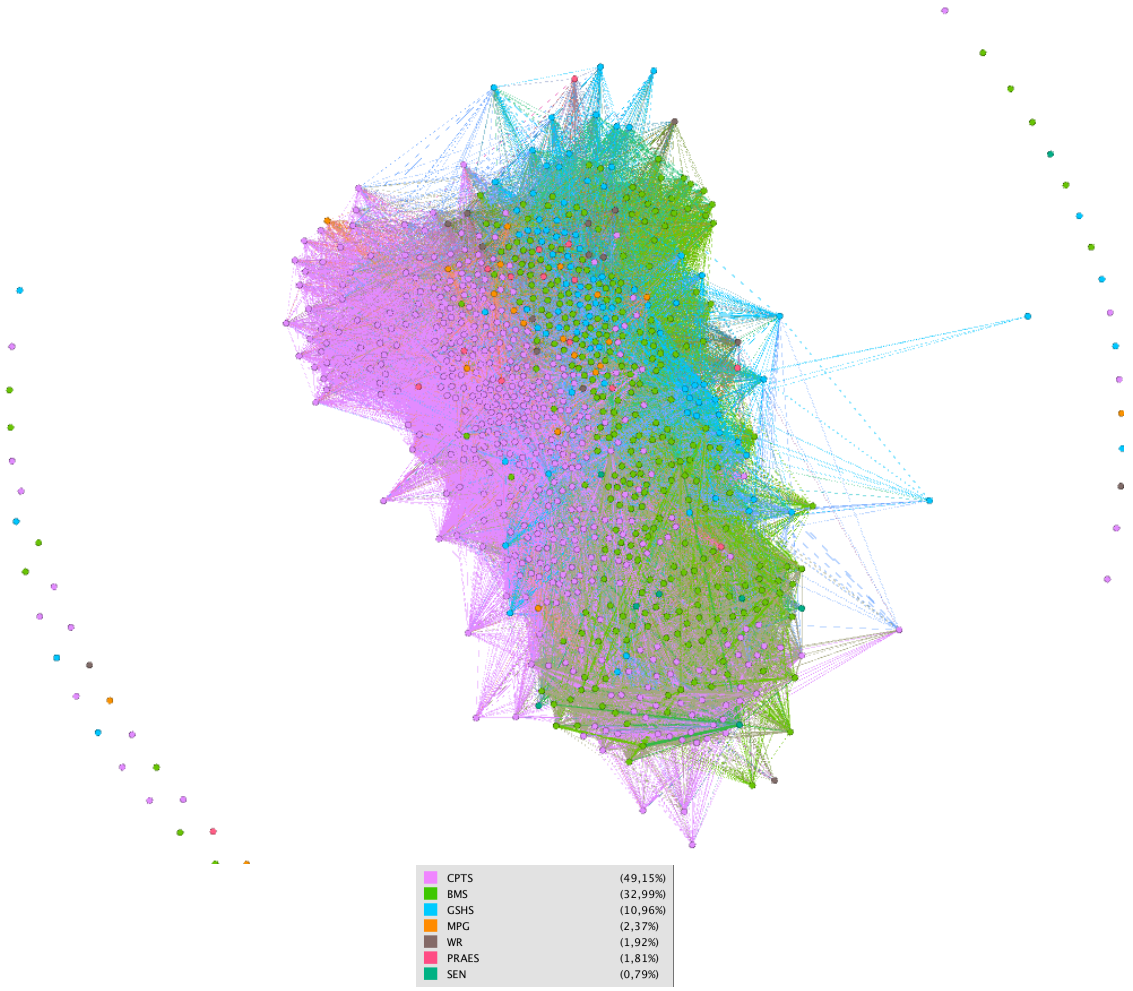


Abbildung 8.55: Koinfluenz, eingefärbt sind die Sektionen der MPG, Darstellung mit Gephi, Fruchtermann-Reingold (Area 10000, Gravity 10.0, Speed 1)

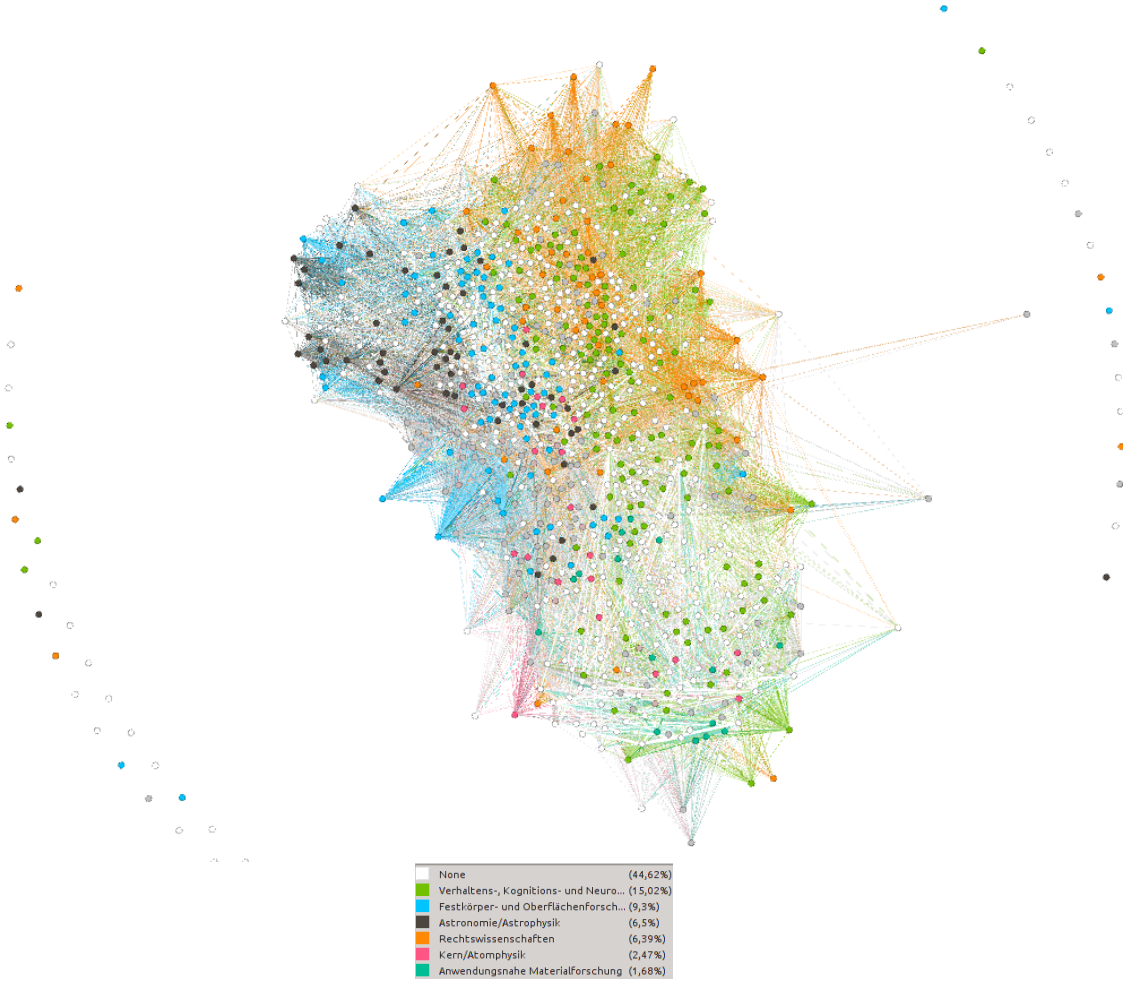


Abbildung 8.56: Koinfluenz, eingefärbt sind wissenschaftlichen Cluster, Darstellung mit Gephi, Fruchtermann-Reingold (Area 10000, Gravity 10.0, Speed 1)

	ber	grue	zukun
abs_betw	0.675991	0.224831	0.099178
abs_close	0.763801	0.141459	0.094740
abs_degree	0.767676	0.151935	0.080390
betw	0.251299	0.450568	0.298133
betw_abw	0.239577	0.474923	0.285501
betw_cnt	0.741520	0.153157	0.105323
close	0.333178	0.332646	0.334176
close_abw	0.280154	0.426514	0.293332
close_cnt	0.767631	0.136483	0.095887
degree	0.343207	0.366175	0.290618
degree_abw	0.270212	0.486158	0.243631
degree_cnt	0.775318	0.146252	0.078431

Abbildung 8.57: Verteilung auf Typen, Mittelwerte, absolute Werte (abs) und Standardabweichung (abw) vom Mittelwert.

	ber	grue	zukun
abs_betw	0.750312	0.136654	0.113034
abs_close	0.758264	0.145923	0.095813
abs_degree	0.748475	0.143324	0.108201
betw	0.312500	0.306819	0.380681
betw_abw	0.290001	0.341398	0.368601
betw_cnt	0.752986	0.142024	0.104990
close	0.326884	0.339118	0.333998
close_abw	0.256575	0.250787	0.492638
close_cnt	0.745214	0.144303	0.110483
degree	0.312379	0.322462	0.365158
degree_abw	0.243496	0.228107	0.528397
degree_cnt	0.745199	0.144438	0.110363

Abbildung 8.58: Verteilung auf Typen Mittelwerte, absolute Werte (Koinfluenz)

Kapitel 9

Modellierung der Datenbank der administrativen Dokumente zum Bau der Kuppel des Doms in Florenz

Die folgende Fallstudie hat ihren Schwerpunkt in Fragen der semantischen Modellierung. Die Herausforderung ist die Überführung einer architekturhistorischen Datenbank in eine *RDF*-basierte Darstellung. Ziele sind hier exemplarische Auswertungen, die mit der bestehenden Datenbank nicht durchzuführen waren. Zusätzlich wollen wir in diesem Beispiel zeigen, wie sich Daten durch die Verbindung mit externen Ressourcen anreichern lassen. Zugleich zeigt dieses Beispiel jedoch auch, welche Probleme noch immer bestehen, Informationen, die als *Linked Open Data* angeboten werden, verlässlich zu Forschungszwecken zu nutzen. Schließlich soll mit der Modellierung basierend auf *CRM* ein Beitrag geleistet werden, diese Daten langfristig zu sichern.

Die Datenbank, die wir betrachten, ist das Ergebnis jahrelanger akribischer Arbeit einer Gruppe von Historikerinnen und Historikern unter der engagierten Leitung von Margaret Haynes, die Bauakten aufzuarbeiten, die aus der Zeit des Baus der Kuppel des Doms von Florenz überliefert sind. Diese Dokumente stellen eine einzigartige Quelle dar, die Aufschlüsse über die technischen und organisatorischen Randbedingungen eines für die damalige Zeit nicht für durchführbar gehaltenen Vorhabens des Baus einer Kuppel mit dem Durchmesser von 45 Metern und einer Höhe von 107 Metern ohne ein stützendes Gerüst, das aufgrund der schieren Menge an benötigtem Holz nicht realisierbar gewesen wäre, gibt.

Die Datenbank hat eine doppelte Struktur, die unter dem Gesichtspunkt eines Modellierungsansatzes von besonderem Interesse und zugleich für die Arbeit mit historischen Quellen paradigmatisch ist. Auf der einen Seite wird detailliert die physische und logische Struktur des Archivs wiedergegeben, auf der anderen Seite wird eine inhaltliche Klassifikation der Daten vorgenommen. Ziel der Modellierung ist es, beide Aspekte abzubilden. Insbesondere wird neben einer Ontologie für die archivalische Beschreibung schrittweise beispielhaft eine Ontologie entwickelt, die die Inhalte der Datenbank zu erschließen hilft.

Struktur und Anlage der Datenbank sind pragmatisch über die Zeit gewachsen. Standards sowohl

für den Einsatz der Software als auch für eingesetzte Formate und Datenmodelle waren zu Beginn der Arbeiten nicht vorhanden oder zumindest nur gering verbreitet. Funktionierende Arbeitsabläufe sollten möglichst nur dann verändert werden, wenn äußerliche Notwendigkeiten dies unumgänglich machen, um die Effektivität der Arbeit nicht zu gefährden. In der Geschichte des Projektes waren es vor allem der Wechsel von Betriebssystemen und Hardware, der Veränderungen erforderte. Derartige Umstiege sind jedoch immer mit Kosten verbunden, die in solchen Projekten in der Regel nicht durch Eigenmittel aufgebracht werden können.

Die Datenbank umfasst das Wissen und die Arbeit von nahezu zwei Generationen von Wissenschaftlerinnen und Wissenschaftlern und stellt in dieser Form ein Kulturgut dar, für das Wege der Erhaltung gefunden werden müssen. Gerechtfertigt wird der Aufwand dafür umso mehr, wenn der Mehrwert der Aufarbeitung der vorhandenen Quellen in digitaler Form auch für die wissenschaftliche Arbeit deutlich gemacht werden kann.

9.1 Derzeitige Realisierung der Datenbank und der Webpräsentation

In der jetzigen Form ist die Datenbank im Internet frei zugänglich. Die nach außen bereitgestellte Version ist im Wesentlichen statisches HTML und durch Indizes und eine komplexe Linkstruktur erschlossen, die im Großen und Ganzen ohne dynamische Elemente auskommt. Die Designentscheidungen wurden zu einem Zeitpunkt getroffen, als der Betrieb dynamischer Webseiten mit der damals am Institut zur Verfügung stehenden Hardware- und Serverausstattung nicht verlässlich möglich war. Zugleich stand der Aspekt einer möglichst soft- und hardwareunabhängigen Implementation im Zentrum der Überlegungen. Dies insbesondere vor dem Hintergrund, dass das Projekt in den späten 1990er Jahren bereits einmal kurz davor stand, einen Großteil der Daten zu verlieren, da die Hardware, auf der die eigens für das Projekt erstellte Datenbanksoftware lief, nicht mehr gewartet werden konnte und der Ausfall drohte.

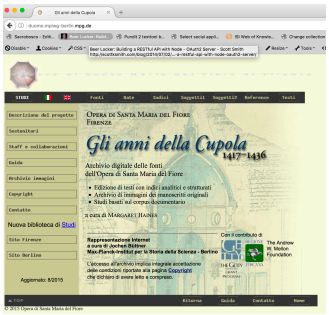


Abbildung 9.1: Ansicht der Einstiegsseite (*http://duomo.mpiwg-berlin.mpg.de*, Stand: 5.8.2017)

Eine Dokumentation, geschweige denn ein lesbares Export- oder Austauschformat für die Daten, lag nicht vor. Jochen Büttner hat am Max-Planck-Institut für Wissenschaftsgeschichte datenarchäologische Arbeit geleistet und die Daten retten können, so dass sie nun in einem neuen System weiter bearbeitet werden können. Als Austausch- und Exportformat dient ein Dump der Datenbank in XML. Zum Zeitpunkt der Arbeit an den Daten waren jedoch XML-Standardformate für archivalische Daten noch wenig verbreitet, *TEI*[1] war für Nichteditionswissenschaftler mangels Softwareverfügbarkeit und Aufarbeitung für Nicht-Spezialisten noch nicht mit vertretbaren Aufwand anwendbar, so dass die verwendete Kodierung in XML keinen externen Standards folgt. Die Struktur der Daten ist jedoch aus dem XML-Code und der

[1] Zu TEI siehe die Ausführungen in Abschnitt 2.5.

Webseite leicht zu erschließen. Mittels Perl-Skripten wird aus diesem XML-Code im wesentliches statisches HTML erzeugt und dieses im Internet präsentiert.

9.1.1 Struktur der Datenbank in der Webdarstellung

Im Zentrum der Datenbank steht die inhaltliche und archivalische Erschließung der einzelnen Teile des Archives. Hierbei werden zunächst die archivalischen Einheiten (*archval unit*,Tabelle 9.1, Abbildung 9.2) beschrieben. Jede archivalische Einheit besteht aus einzelnen Folios,[2] die in der üblichen

Signatur	Signatur
Descriptive Title	Beschreibender Titel der Editoren
Original Title	Titel wie er im Archiv vorgefunden wurde (italienisch)
Terminal dates	Zeitraum, in dem die Einträge erstellt wurden
Physical description	Beschreibung der Archivalie
Conservation	Erhaltungszustand der Archivalie
Language	Sprache der Einträge
Person(s) accountable	Verantwortliche Personen
Scribe	Schreiber der Einträge
Contents	Kurze inhaltliche Zusammenfassung
Reproduction	Reproduziertes Material
Edition	Vorhandene Editionen
Annnotations	Vorhandene Annotationen im Dokument

Tabelle 9.1: Felder zur Beschreibung der Archivalischen Einheit (Archival Unit)

Weise durchgezählt werden. Jedes Folio hat eine Identifikationsnummer. Vorder- und Rückseite werden mit v (*verso*) und r (*recto*) gekennzeichnet. Logisch zusammenhängende Einträge auf den Seiten werden in der Webdarstellung als Dokumente (*documents*) bezeichnet. Auf einer Seite können mehrere Dokumente verzeichnet sein (Abbildung 9.3). Die Webdarstellung gibt jeweils die Möglichkeit, auf die digitalisierten Seiten zurückzugreifen. Abbildung 9.4 zeigt drei als unterschiedliche Dokumente identifizierte Teile einer Seite durch Rahmen hervorgehoben.

Für jedes einzelne Dokument werden zunächst die wesentlichen Daten und der Text (*Transcription of text and essential data*) dargestellt (Abbildung 9.3,Tabelle 9.3). Im zweiten Teil eines jeden Eintrages finden sich analytische Daten, die sich aus der näheren Interpretation des Inhaltes ergeben. Dies sind zunächst erwähnte Personen und Institutionen mit ihren Rollen, wie sie in den Dokumenten aufgefunden werden (siehe Tabelle 9.4).

Die komplexeste Struktur steht hinter den unter *guided research* angebotenen Informationen. Hier wird der Inhalt erschlossen und Handlungen werden unter Angabe der betroffenen Objekte, Materialien, der beteiligten Personen und Institutionen systematisch aufbereitet (siehe Tabelle 9.2). Dabei werden vorgegebene analytische Kategorien verwendet, die die für den Bau und dessen Arbeitsorganisation wesentlichen Strukturen abbilden. Diese Kategorien ermöglichen eine Vernetzung der

[2] d.h. einzelnen Seiten

Personnel	Handlung in die Personen involviert sind; erfasst wird eine Klassifikation der Handlung, sowie die Personen oder Institutionen die an dieser Handlung beteiligt sind, sowie eine Kurzbeschreibung der Handlung.
Destinations	Ort auf den sich die Handlungen in den Einträgen beziehen, z.B. ein Teil des Doms.
Objects	Klassifikation der Objekte, die erwähnt werden sowie eine nähere Beschreibung
Materials	Klassifikation der erwähnten Materialien sowie eine nähere Beschreibung
Iconography	Klassifikation und Beschreibung

Tabelle 9.2: Handlungen und Abläufe

Abbildung 9.2: Ansicht einer Beschreibung einer archivalischen Einheit.
(*http://duomo.mpiwg-berlin.mpg.de/ENG/AR/ARS007.HTM*, Stand: 5.8.2017)

unterschiedlichen Einträge. Sie sind jeweils mit Übersichtsseiten verlinkt, auf denen alle Dokumente zu einem bestimmten Aspekt verzeichnet sind. Abbildung 9.6 zeigt einen Ausschnitt der Seite zu *gabelles – new* (Neue Abgaben). Letzter Aspekt ist schließlich die chronologische Einordnung der

Abläufe, der im Dokument dokumentierten Ereignisse (*Chronological*).

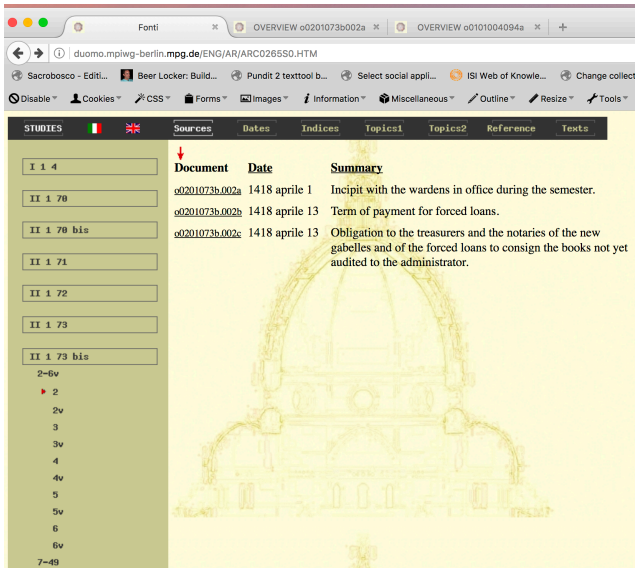


Abbildung 9.3: Drei Dokumente auf einer einzelnen Seite. (*http://duomo.mpiwg-berlin.mpg.de/ENG/AR/ARC0265S0.HTM*, Stand: 5.8.2017)

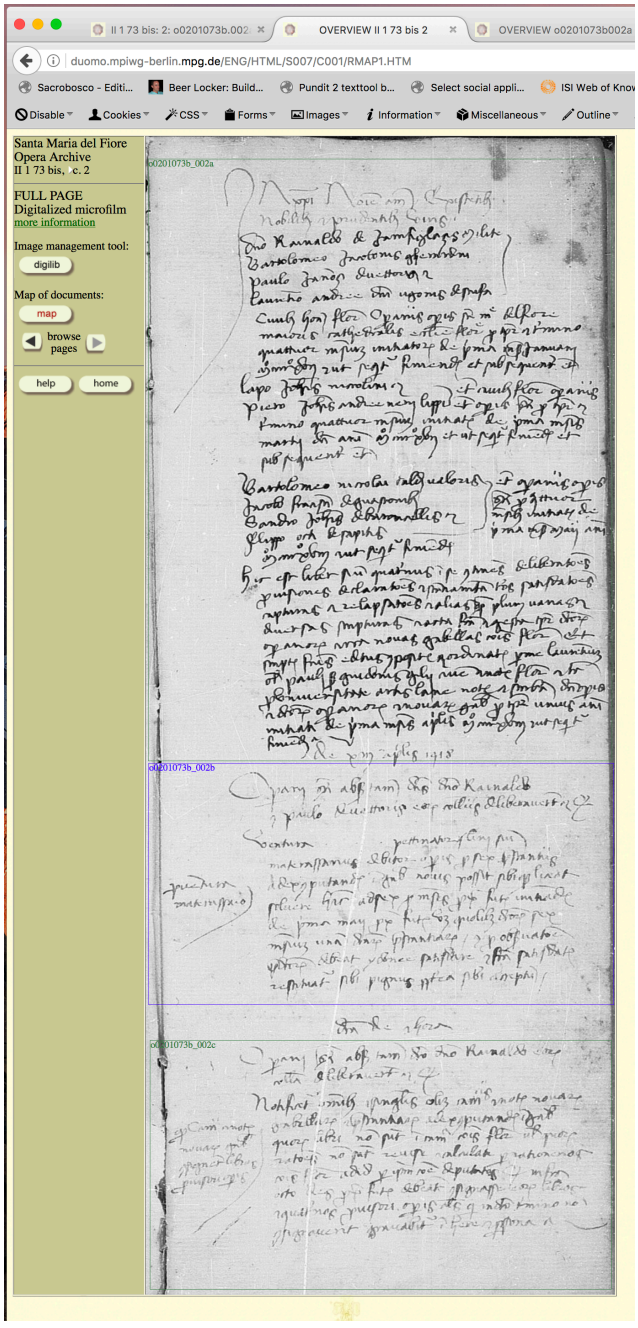


Abbildung 9.4: Drei Dokumente auf einer einzelnen Seite. (*http://duomo.mpiwg-berlin.mpg.de/ENG/HTML/S007/C001/RMAP1.HTM*, Stand: 5.8.2017)

Abbildung 9.5: Ansicht eines Dokumentes
(*http://duomo.mpiwg-berlin.mpg.de/ENG/CA/CA200001650.HTM*, Stand: 5.8.2017)

Document	ID des Dokuments
Date	Datum des Eintrages
Summary	Kurze Zusammenfassung des Inhaltes (EN,IT)
Source	Beschreibung der Quelle bestehend aus Signatur und Angabe der Seiten
Title	Originaltitel des Eintrages (IT,LA)
Text	Transkription des Textes (mit Verweisen auf Anmerkungen in den Notizen)
Notes	Anmerkungen zur Transkription
Bibliography	Verweis auf Editionen oder andere Sekundärquellen, in denen der Text erwähnt wird.

Tabelle 9.3: Deskriptive Felder zu den Einzeldokumenten

Names and Roles	Erwähnte Personen und Rolle, die sie in dem Eintrag haben.
Institutions	Erwähnte Institutionen

Tabelle 9.4: Personen und Institutionen

Abbildung 9.6: Ausschnitt der Einträge zu *gabelles - new* (Neue Abgaben) (*http://duomo.mpiwg-berlin.mpg.de/ENG/CA/CA200016S0.HTM*, Stand: 5.8.2017)

9.1.2 Die XML-Darstellung und ihre Analyse

Der für die Webdarstellung benutzte Export der Datenbank ist auch der Ausgangspunkt für die hier im Weiteren dargestellte Modellierung in OWL. Die Struktur lässt sich gut aus dem dazu gehörigen *relax-ng*-Schema ablesen.[3]

[3] Siehe [268] und dort genauer [270]. Ausschnitte des Schemas sind in den Abbildungen 9.7,9.8, 9.10, 9.11 und 9.9 zu sehen.

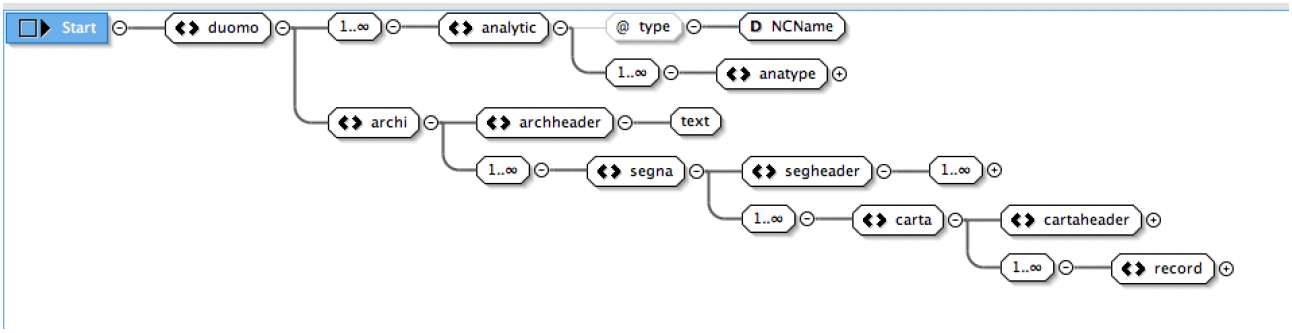


Abbildung 9.7: Struktur der Datenbank; relax-NG Schema; Ausschnitt mit segna (Signatur), carta (Karte), record (Eintrag), erstellt mit Oxygen [180].

Das Schema zerfällt in zwei Hauptteile: einerseits die analytischen Kategorien, die innerhalb des Tags *<analytic>* beschrieben werden, andererseits die Darstellung der Ordnung des Archivs selbst mit einer strikten Hierarchie von Archiv *<archi>*, Signatur *<segna>*[4] und Karten *<carta>*[5]. Auf diesen Karten sind wiederum Einträge erfasst – also die eigentlichen dokumentierten Ereignisse. Hierbei tauchen strukturelle Probleme auf, da ein Eintrag[6] sich über mehrere Karten erstrecken kann. Dieses ist jedoch in dem nach der physikalischen Struktur hierarchisch aufgebauten XML nicht direkt darstellbar, da sich in diesem Falle überlappende Tags ergeben hätten. Die Unterordnung der Einträge, wie sie das Modell durch die Unterordnung des *<record>* unter die *<carta>* suggeriert, ist in dieser Form inhaltlich nicht zutreffend und wird umgangen, indem ein Eintrag ein zusätzliches Attribut *end_on_carta* in *<record>* enthält. Die analytische Beschreibung der Einträge erfolgt nun in den *<record>* untergeordneten Tags *<campianalitici>*, *<identificator>* und *<indice>*, hinzu kommen Verweise auf die Literatur in *<reference>* und administrative Daten.

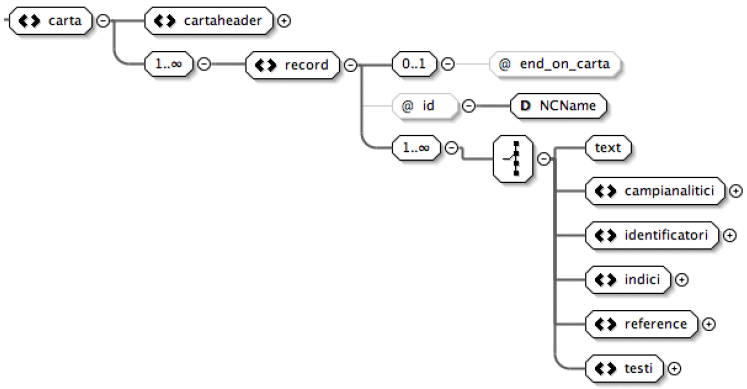


Abbildung 9.8: Struktur der Datenbank; relax-NG Schema; Ausschnitt mit der Struktur für einen *<record>*; erstellt mit Oxygen.

[4]Diese entsprechenden archivalischen Einheiten in der Web-Darstellung.
[5]Diese entsprechen Folioseiten.
[6]In der Webdarstellung entspricht ein Record einem Dokument.

Name auf englischer Webseite	Beschreibung	xpath container	
		archi/segna	
Signatur			**segheader**
Descriptive Title	Beschreibender Titel der Editoren		
Original Title	Titel wie er im Archiv vorgefunden wurde (italienisch)		
Terminal dates	Zeitraum in denen die Einträge erstellt wurden		
Physical description	Beschreibung der Archivalie		
Conservation	Erhaltungszustand der Archivalie		
Language	Sprache der Einträge		
Person(s) accountable	verantwortliche Personen		
Scribe	Schreiber der Einträge		
Contents	Kurze Inhaltliche Zusammenfassung		
Reproduction	Reproduziertes Material		
Edition	Vorhandene Editionen		
Annnotations	Vorhandene Annotationen im Dokument		
		archi/segna/carta	
Document	ID des Rekords		record/@id
			record/@end_on_carta
chronological	Zeitraum des Events das im Record beschrieben wird		record/identificatori/datrf/startdate,record/identificatori/datrf/endddate
Date	Datum des Eintrages		record/identificatori/datdc/startdate,record/identificatori/datdc/endddate
Summary	Kurze Zusammenfassung des Inhaltes (EN,IT)		record/identificatori/reges/english; reges/italian
	Textblock ID		record/identificatori/textblockid
	Typisierung		record/identificatori/tipol
Source	Beschreibung der Quelle bestehend aus Signatur und Angabe der Seiten		cartaheader/cartanr
	(Angabe der Seiten)		cartaheader/cartaant
Title	Originaltitel des Eintrages (IT,LA??)		record/testi/titol
Text	Tanscription des Textes (mit Verweisen auf Anmerkungen in den Notizen)		record/testi/testo
Notes	Anmerkungen zur Transkription		record/testi/testo/node
Administrative Informtionen			record/reference/trasc;record/reference/revis
Bibliography	Verweis auf Editionen oder andere Sekundärquellen in denen der Text erwähnt wird.		record/reference/bibli
Names and Roles	Erwähnte Personen und ihre Rolle, die sie in dem Eintrag spielen. Personen können auch in Gruppen zusammen gefasst werden,		record/indici/nomiq/(name,role)
groups	Gruppen können gruppen enthalten		record/indici/group
	Erwähnte Rollen ohnen namen		record/indici/nomiq/role
Institutions	Erwähnte Institutionen		record/indici/isit
	Handlung in die Personen involviert sind, erfasst wird jedenwalls eine Klassifikation der Handlung, sowie die Personen oder Instituionen die an dieser Handlung beteiligt sind, sowie eine Kurzbeschreibung der Handlung.		
Personnel		record	campianalitici/PERSONALE
		record	campianalitici/PERSONALE/type/ptr/@target
			campianalitici/PERSONALE/type/freetext
Destinations	Ort auf den sich die Handlungen in den Einträgen beziehen, z.B. ein Teil des Doms.		campianalitici/DESTINAZIONI
Objects	Klassifikation der Objekte, die erwähnt werden, sowie eine nähere Beschreibung		campianalitici/OGETTI
Materials	Klassifikation der erwähnten Materialien, sowie eine nähere Beschreibung		campianalitici/MATERIALI
Iconography	Klassikation und Beschreibung		campianalitici/ICONOGRAFIA

Abbildung 9.9: Webdarstellung und XML-Struktur

9.1.3 Analytische Beschreibung

Jedem <*record*> werden eine oder mehrere analytische Kategorien bzw. Subkategorien zugewiesen. Diese Zuordnungen finden sich innerhalb des Containers <*campianalitici*> gruppiert nach 11 Oberkategorien (Abbildung 9.11).

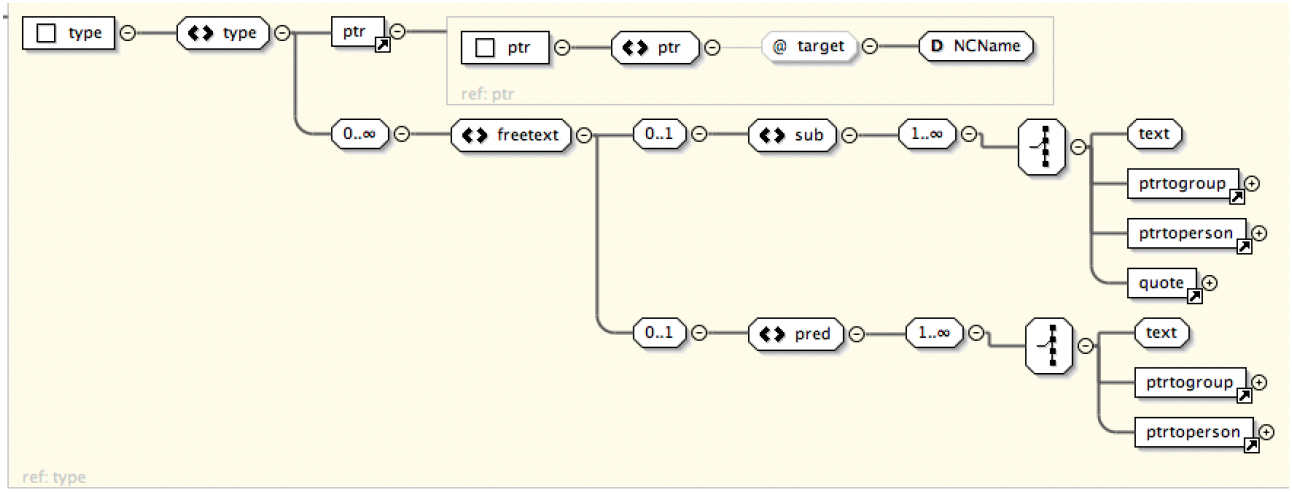


Abbildung 9.10: Struktur der Datenbank; relax-NG Schema; Ausschnitt mit der Struktur von <*type*>; erstellt mit Oxygen.

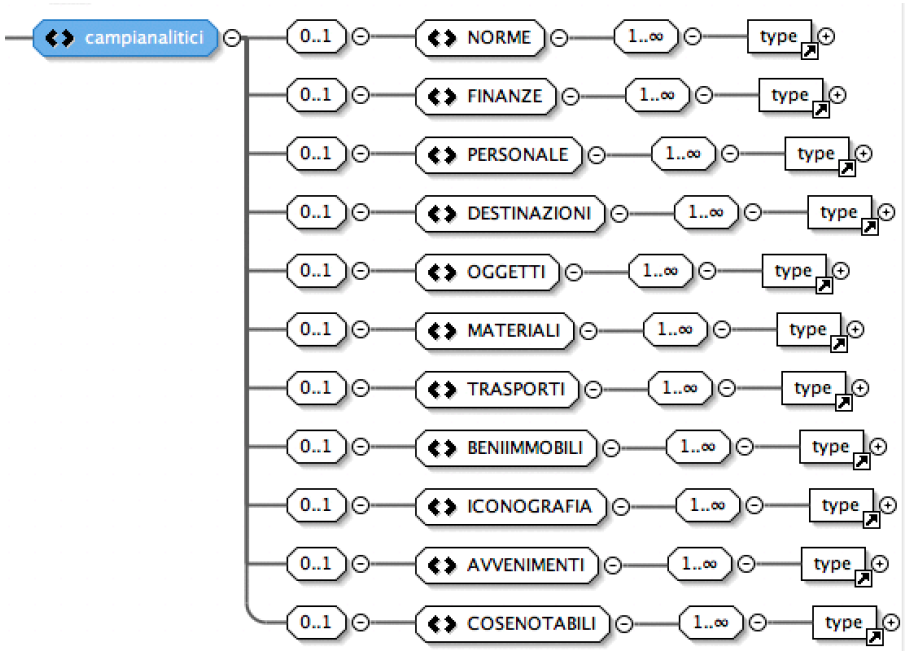


Abbildung 9.11: Struktur der Datenbank; relax-NG Schema; Ausschnitt mit der Struktur für die analytischen Kategorien; erstellt mit Oxygen.

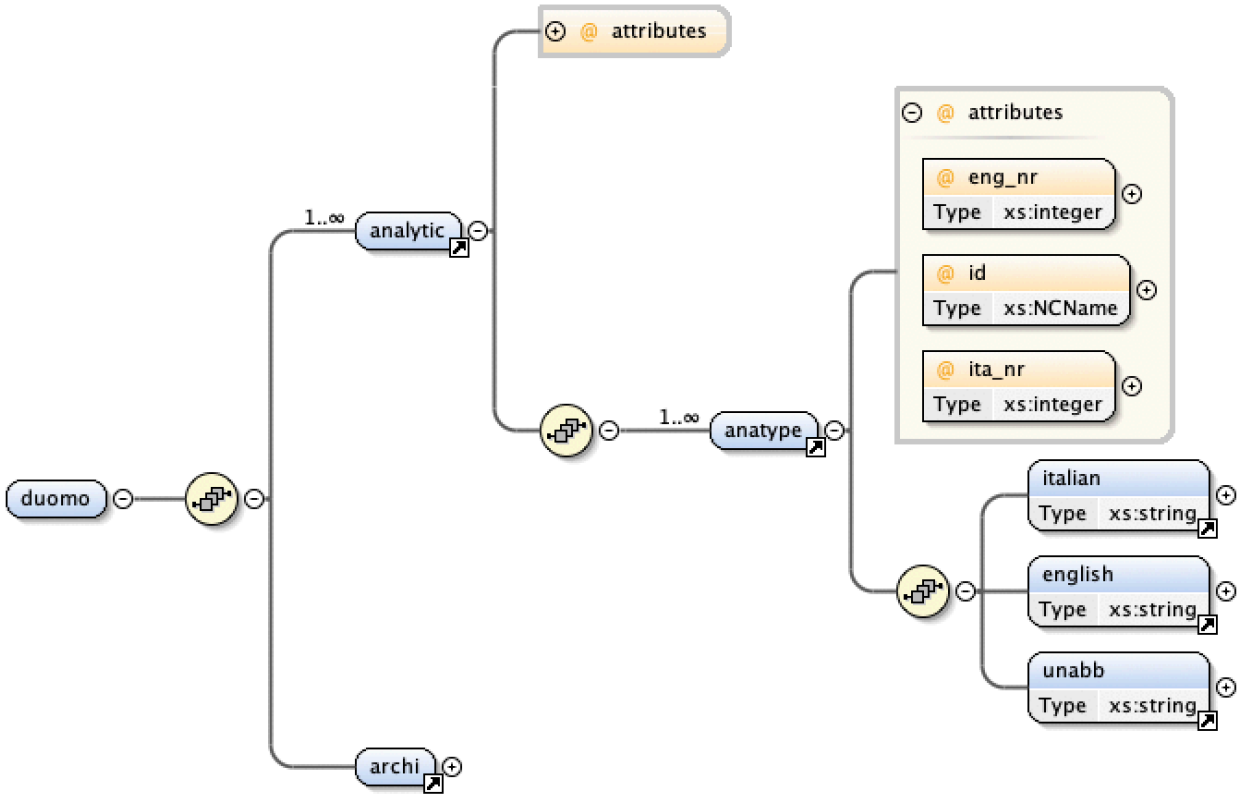


Abbildung 9.12: Struktur der Datenbank; relax-NG Schema; Struktur von <*analytic*>; erstellt mit Oxygen.

<*type*> (Abbildung 9.10) enthält immer einen Pointer <*ptr*> auf eine analytische Kategorie sowie optional eine oder mehrere Freitextbeschreibungen <*freetext*>, die die Zuordnung näher beschreiben und selber Verweise auf Personen <*ptrtoperson*> oder Personengruppen <*ptrgrp*> beinhalten können. Die analytischen Kategorien sind mit <*analytic*> annotiert und haben jeweils Unterkategorien <*anatype*> (Abbildung 9.12).

Während mit <*ptr*> auf Kategorien <*anatype*> verwiesen wird, die für alle <*record*> gleich sind, verweisen <*ptrtoperson*> und <*prtgrp*> immer nur auf Personenindizes innerhalb eines Record (siehe 9.13). Es existiert also in der XML-Darstellung kein globaler Personenindex. Erwähnte Personen, Rollen und Institutionen sind jeweils nur in den einzelnen Einträgen in <*indici*> zusammengefasst. Eine Transkription des Titels (<*titol*>) jedes Eintrages und des Eintrages selbst ist in der XML-Struktur enthalten (<*testi*>). In den Volltexten (<*testo*>) sind jeweils nur Layoutauszeichnungen, aber keine semantischen Auszeichnungen vorgenommen worden.

9.2 Modellierung in FRBRoo

Ausgehend von der XML-Darstellung beginnen wir nun mit der Umsetzung des geschilderten Datenbankschemas in FRBRoo.[7] Zunächst gilt es, sich Klarheit über die einzelnen Entitäten zu verschaffen, die beschrieben werden sollen. Die grundsätzliche Frage, die vor jedem Modellierungsansatz zu

[7] Es wird nur in sehr geringem Maße auf neue Elemente aus FRBRoo zurückgegriffen.

```
<duomo>
...
<analytic type="destinazioni">
...
<anatype id="DESTINAZIONI03" ita_nr="460" eng_nr="460">
<italian>Duomo - cappelle</italian>
<english>Duomo - chapels</english>
<unabb>Duomo - cappelle</unabb>
</anatype>
...
</analytic>
...
<record>
...
<indici>
<nomiq id="o0101004.095a01"><name>Lorenzo di Bartoluccio</name><role>orafo</role></nomiq>
...
</indici>
<campianalitici>
<PERSONALE>
<type><ptr target="PERSONALE10"/>
<freetext><sub><ptrtoperson target="o0101004.095a01"/> </sub>
  <pred> salario per sepoltura San Zanobi</pred>
  </freetext>
</type>
</PERSONALE>
<DESTINAZIONI>
<type><ptr target="DESTINAZIONI03"/>
<freetext><sub>cappella di San Zanobi, cassa</sub></freetext>
</type>
</DESTINAZIONI>
....
</campianalitici>
...
</record>
```

Abbildung 9.13: Das Beispiel für einen Container <*campianalitici*> mit mit Pointern.

beantworten ist, lautet: Was soll beschrieben werden? In unserem Falle ist dies eine elektronische Repräsentation des Archivs in Form einer Repräsentation der Dokumente als digitale Faksimiles auf der einen Seite und auf der anderen Seite eine inhaltliche Interpretation des Inhaltes dieses Archivs. Das wesentliche Objekt, das es zu beschreiben gilt, ist hierbei der einzelne Eintrag, der auf einem Teil einer Seite oder auch auf mehreren Seiten der archivalischen Einheiten auffindbar ist.

Wir teilen die beiden genannten Aspekte zunächst auf. Zunächst entwickeln wir ein Modell für die digitale Repräsentation und archivalische Beschreibung der Dokumente sowie die archivalische Einheit; im zweiten Schritt wird dann ein Modell für die inhaltliche Analyse entwickelt. Die Trennung folgt hierbei der von den Autorinnen der Datenbank vorgegebenen Aufteilung. Bei der inhaltlichen Analyse komme ich darauf noch einmal zurück.

Die Archivstruktur bilden wir auf die in Abbildung 9.15 dargestellte Struktur ab.[8] Zur Vereinfachung der Lesbarkeit gilt im folgenden immer, dass alle Klassen bzw. Properties, die mit E[0-9]+/P[0-9]+ beginnen, zu CRM, diejenigen mit F[0-9]+/R[0-9]+ zu FRBroo und alle anderen zur neuen Duomo-Ontologie gehören, wenn nicht explizit ein Namespace angegeben wird.

[8] Die detailierte Aufschlüsselung findet sich in [267].

```
<testi language="volgare">
<testo><P indent="yes">E<note><P indent="no"> Nel margine superiore, circa a met&#224; rigo,
compare di mano pi&#249; tarda,
segnalata da Poggi come quella di Carlo Strozzi, il numero <quote>1388</quote></P></note>
detti operai imprima alluogano al detto Lorenzo la chassa in che &#224; ' s
tare<note><P indent="no"> Segue depennato <quote>nell'altare dell-</quote>.</P></note>
il corpo del glorioso Sam Zanobi nell'altare della chappella di Sancta Maria del
Fiore al decto sancto diputata di lung<italics>h</italics>eza braccia 3 il sodo e pi&#249;
quello gitteranno le cornice, le quali<note><P indent="no"> Segue depennato
<quote>far&#224; secondo in decta</quote>.</P></note> debba fare secondo parr&#224;
a<connect/>llui, larga quanto si conviene<note><P indent="no"> Segue depennato <quote>al che</quote>.
</P></note> a l'equa
distanzia intorno drento alle cornici, alta braccia 1 e 1/2 col
coperchio che sia alla forma del modello, storiata intorno intorno
delle storie di San Zanobi, secondo diranno e due diputati
sopra a<connect/>cc<italics>i</italics>&#242;.
Tutta la sopradetta cassa &#224; essere d'ottone dorato,
d'arienti comessi di proferiti o d'altre materie, chome parr&#224;
a' detti due diputati sopra<connect/>cc<italics>i</italics>&#242;.</P>
<P indent="yes">E debela lavorare diligentemente e bene quanto pi&#249; si pu&#242; a chiarig
<italics>i</italics>one degli operai di Sancta Maria del Fiore che pe' tempi saranno; s&#236;
veramente che detta cassa con tutti i sua imbasamenti non passi libbre 5.000 e passandola sia
quel pi&#249; a danno del detto maestro.</P><P indent="yes">E pi&#249; impromette
il detto Lorenzo rendere compiuta la detta sepoltura di qualunche cosa a' detti operai
infra anni 4.<note><P indent="no"> Il numero <quote>4</quote> &#232; scritto, forse in
un secondo tempo, sulla parte iniziale di uno spazio pi&#249; ampio <insertion>di circa
8 lettere</insertion> rimasto in bianco.</P></note> E se gli passasi, s'intenda il d&#236;
del finito tempo non potere pi&#249; lavorare in su detta sepoltura e sia lecito agli
operai che e que'<note><P indent="no"> <quote>e que'</quote> aggiunto nell'interlinea,
a sostituzione di <quote>pe'</quote>, depennato.</P></note> tempi fiano poterlla allogare
la detta sepoltura a qualunche altra persona a<connect/>lloro piacer&#224;.</P>
<P indent="yes">E pi&#249; e detti operai impromettono al detto Lorenzo dare ottone,
pietre, ariento e oro e ogni altra materia che entrasse in detta chassa, chome diranno e
detti diputati sopra<connect/>cc<italics>i</italics>&#242; a que' tempi sarano pel detto L
orenzo richieste.</P><P indent="yes">E pi&#249; impromettono i detti operai dare o fare
[...]
</testi>
```

Abbildung 9.14: Beispiel für eine Textauszeichnung

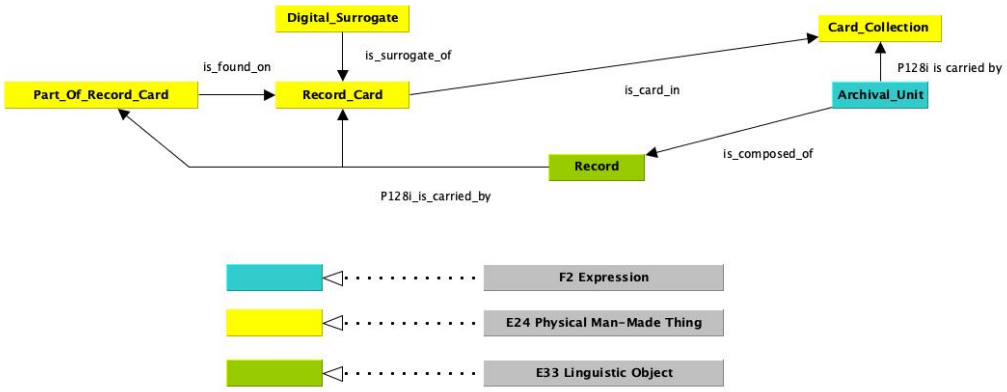


Abbildung 9.15: Übersicht über die Modellierung des Archivs

9.2.1 Erstellung einer archivalischen Einheit im Archiv

In der ereignisorientierten Formulierung von CRM beschreiben die Metadaten das Ereignis der Erstellung einer einzelnen archivalischen Einheit (Tabelle 9.1). In den Metadaten liegen uns Informationen über den Schreiber (**duomo:Scribe**),[9] den Entstehungszeitpunkt des Eintrages sowie den Originaltitel vor. Es ergibt sich damit die in Abbildung 9.16 dargestellte Situation. Zusätzlich liegen jeweils

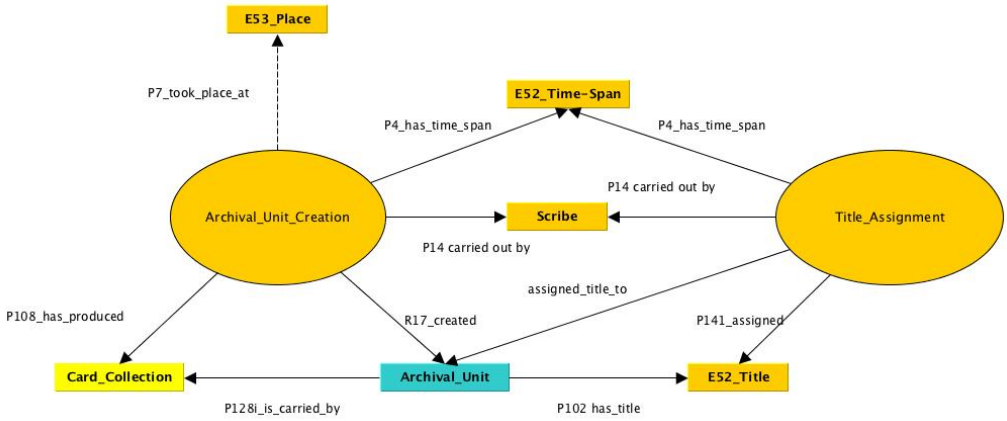


Abbildung 9.16: Modellierung der Erstellungssituation durch den Schreiber

noch ein Titel, der von den Autorinnen der Datenbank der Einheit zugewiesen wurde, eine kurze Inhaltsangabe, sowie Personen, die in der archivalischen Einheit als verantwortlich für diese Einheit aufgefunden wurden, vor. Graphisch ist diese Situation in Abbildung 9.17 beschrieben.[10] Damit sind die wesentlichen Felder von 9.1 abgedeckt. Die dort fehlenden Felder lassen sich in analoger Weise modellieren. Wir wenden uns nun den einzelnen Einträgen (**duomo:Record**) zu. Diese stellen für die dann im Folgenden behandelte nähere inhaltliche Analyse den eigentlichen Kern dar.

[9] Wir folgen der in Abschnitt 1.5 beschriebenen Notation.
[10] Die detaillierte Beschreibung der Klassen findet sich in [267].

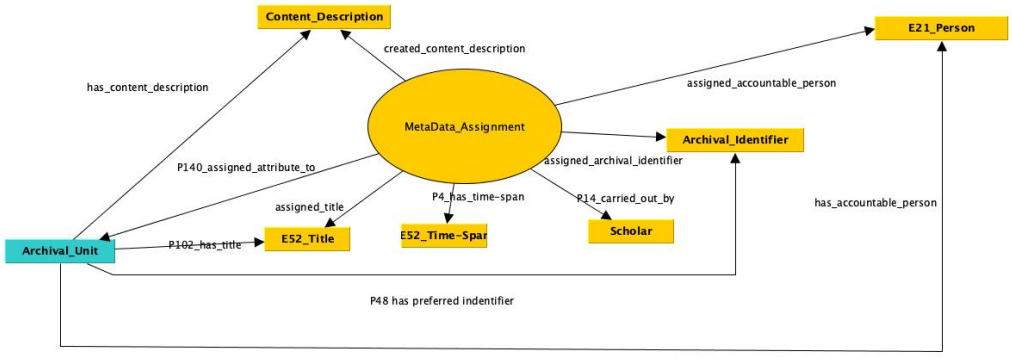


Abbildung 9.17: Beschreibung der archivalischen Einheit durch die Bearbeiter der Datenbank

9.2.2 Beschreibung eines einzelnen Eintrags (Record)

Wir teilen die Modellierungsschritte in eine archivalisch-deskriptive (Felder in Tabelle 9.3) und eine inhaltliche Analyse (Felder in den Tabellen 9.4 und 9.2) auf. Dieses folgt der Aufteilung, wie sie in der Webdarstellung gemacht wurde. Schauen wir auf Tabelle 9.3, so finden wir dort eine Reihe von Werten, die von den Bearbeitern jedes Dokumentes erhoben wurden. Wir nehmen wie bei der Beschreibung der archivalischen Einheit an, dass diese Daten alle bei einem Ereignis angefügt wurden.[11] Damit ergibt sich für die Dokumente (Abbildung 9.18) eine ähnliche Struktur wie zuvor in Abbildung 9.17 für die archivalischen Einheiten gezeigt.

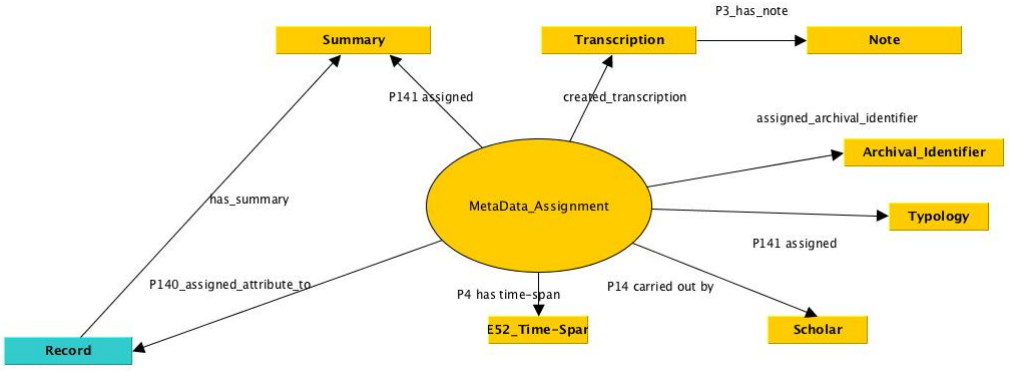


Abbildung 9.18: Beschreibung des Record durch die Bearbeiter der Datenbank

Analog stellen wir die Angaben, die vom eigentlichen Schreiber gemacht wurden, durch eine Unterklasse von **F28_Expression Creation** für den Eintrag dar.

[11] Zumindest liegt in der Datenbank jeweils nur eine Datierung für den gesamten Eintrag vor.

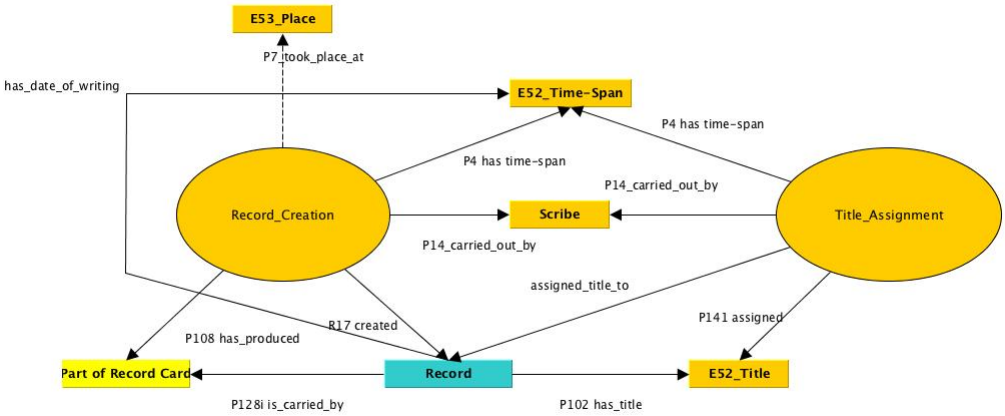


Abbildung 9.19: Modellierung der Erstellungssituation durch den Schreiber

Die eigentlichen Transkriptionen in <*testo*> (siehe Abbildung 9.14) werden nicht direkt in RDF übernommen. Es werden lediglich Verweise auf diese erstellt und jede einzelne Transkription wird im Dateisystem abgelegt.

9.2.3 Modellierung der inhaltlichen Analyse

Wir wenden uns nun der Modellierung der inhaltlichen Analyse – d.h. den Werten in den Tabellen 9.2 und 9.4 – zu. Die Werte dazu finden sich in den Ausgangsdaten innerhalb von <*campianalitici*>, die in Abschnitt 9.1.3 beschrieben wurden. Strukturell wird hier einem Eintrag jeweils ein Ereignis oder mehrere Ereignisse im Kontext der Baustelle durch die Bearbeiter der Karteikarte entsprechend zugeschrieben. Die Beschreibung erfolgt wieder in Form einer Unterklasse von **E65_Creation**.

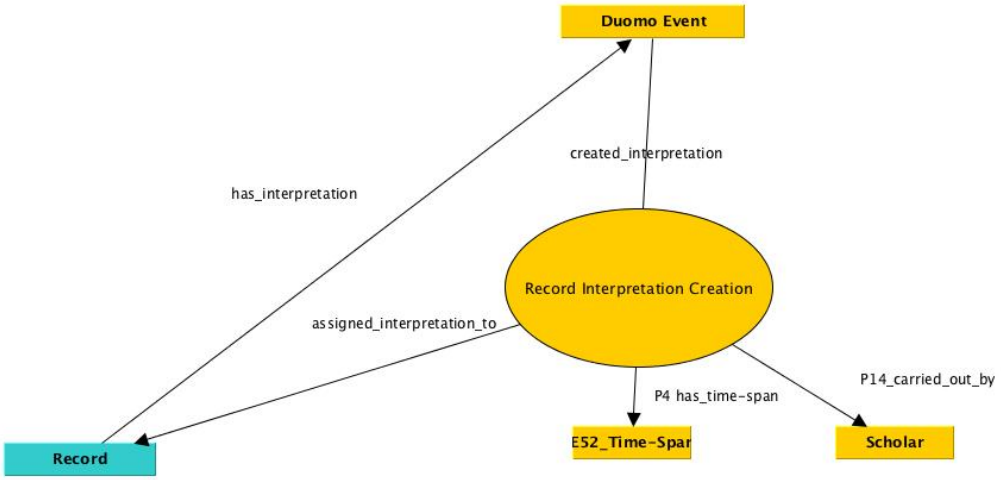


Abbildung 9.20: Modellierung von **duomo:Duomo_Event** durch die Bearbeiter der Datenbank

Die zugewiesenen Ereignisse sind Instanzen einer Klasse **duomo:Duomo_Event**, die selbst eine Unterklasse von **E5_Event** ist. Den Vorgang selbst verstehen wir als eine wissenschaftliche Interpre-

tation der Ereignisse.[12]

9.2.4 Duomo_Event

duomo:Duomo_Event bildet den Kern der inhaltlichen Interpretation der Karteikarten und damit den Ausgangspunkt für alle weiteren Betrachtungen, die sich auf die Analyse der Vorgänge auf der Baustelle beziehen. Im ersten Schritt unserer Modellierung nehmen wir keine weitere Interpretation vor, sondern übernehmen so direkt wie möglich die in der Datenbank vorgefundene Struktur (siehe Abschnitt 9.1.3).

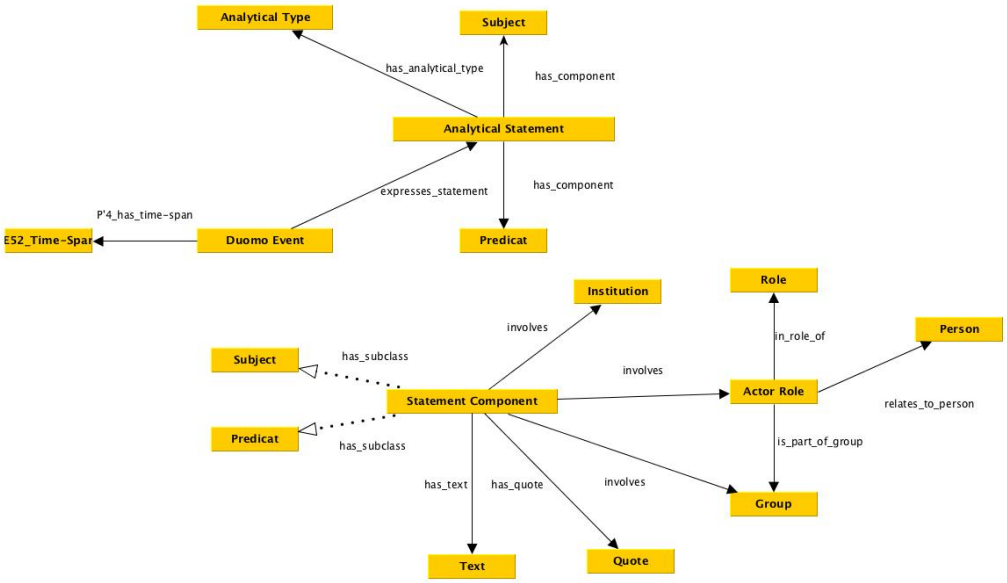


Abbildung 9.21: Struktur von **duomo:Duomo_Event**

Wir fassen hierbei die in der Datenbank für den Inhalt gemachten Aussagen jeweils als **duomo:Analytical_Statement** auf. Die in *<type>* innerhalb von *<campianalitici>* vorgenommene Typisierung wird hierbei zur **duomo:Analytical_Type**, einer Unterklasse von **E55_Type**. Die analytischen Kategorien selbst werden in einem Thesaurus mit zwei Hierarchieebenen in *SKOS* ausgedrückt.[13] Ein Statement besteht aus Aussagekomponenten (**duomo:Statement_Component**), innerhalb derer auf Personen und Institutionen mit entsprechenden Rollen verwiesen werden.

9.2.5 Von der Repräsentation in XML zu Darstellung in RDF

Die eigentliche Transformation des XML-Codes zu einer Darstellung in RDF erfolgt nun mit einem Skript *lex2rdf.py*. Die Ausgabe dieses Skripts besteht aus einer Datei, die die Tripeldarstellung in RDF umfasst. Die einzelnen Transkriptionen werden in einem Ordner als jeweils einzelne Datei abgelegt;

[12] **duomo:Duomo_Event** ist daher auch eine Unterklasse von **duomo:Interpretation**.

[13] Die Kategorien werden mittels des Python-Scripts (*lex2skos.py*) in ein SKOS Vokabular übertragen [272] und liegen im Triplestore im Graphen *duomo:graphs/analyticalclasses*.

auf diese wird dann in RDF nur verwiesen. Von nun an geschehen alle weiteren Schritte auf der Grundlage der RDF-Darstellung.[14]

9.3 Anreicherungen und Analysen

Die Daten im Triplestore werden jetzt schrittweise mit ergänzenden Metadaten angereichert und weiter strukturiert.

9.3.1 Datumsformat

Alle Datumsangaben liegen in den Ausgangsdaten nicht in einem maschinenlesbaren Datumsformat vor. Daher fügen wir zunächst mittels *createDate.py* zusätzliche Informationen über das Datum hinzu. Alle Instanzen von **duomo:fixedDate** erhalten hierbei zusätzliche Properties:

- **sr:has_XSD_date**
- **sr:has_date_qualifier**
- **sr:is_uncertain**

Hierbei hat **sr:_has_date_qualifier** die Werte *ante* oder *post* und beschreibt, ob ein Datum eine obere und untere Grenze eines Zeitintervalles ist. **sr:is_uncertain** enthält Werte laut Tabelle 9.5 für die jeweils unsicheren Anteile eines Datums. Kombinationen ergeben sich aus binärer Addition der Werte. Die so errechneten Zuweisungen werden in einem Graphen[15] abgelegt.

0	Datum sicher
1	Jahr unsicher
2	Monat unsicher
4	Tag unsicher

Tabelle 9.5: Werte für **is_uncertain**; Kombinationen ergeben sich aus binärer Addition der Werte.

9.3.2 Eine erste Visualisierung der Daten über Zeit

Mittels der Visualisierungskomponenten von *metaphactory*[16] lassen sich unmittelbar über Abfragen des Triplestore erste Visualisierungen der Daten erstellen (Abbildung 9.23). So ergibt die Abfrage in Abbildung 9.22 aller Startdaten von Ereignissen ohne größeren Programmieraufwand eine Darstellung der zeitlichen Entwicklung.

[14] Der Import in den Triplestore erfolgt mit *load <file:///usr/local/blazegraph/duomo/out.rdf> into graph duomo:graphs/mainData* in den Graphen *duomo:graphs/mainData*.

[15] *duomo:graphs/datesCalculated*

[16] Siehe Abschnitt 6.6.

```
SELECT ?year (COUNT(distinct ?int) as ?int_n) WHERE {
 ?int rdf:type duomo:Duomo_Event;
    ecrm:P4_has_time-span ?ts.
 optional {
 ?ts duomo:has_XSD_date ?date1
 }
 optional {
 ?ts duomo:has_startDate/duomo:has_XSD_date ?date2
 }
 bind( if(bound(?date1),?date1,?date2) as ?date)
  bind( year(?date) as ?year)
filter (?year < 1900)
 }
 GROUP BY ?year
 ORDER BY ?year
```

Abbildung 9.22: Abfrage nach der Anzahl von Duomo Events pro Jahr (siehe auch im Anhang 12.5)

Analog erhalten wir eine Übersicht der Anzahl der erstellten Records pro Jahr durch Austausch von

```
?int rdf:type duomo:Duomo_Event;
 ecrm:P4_has_time-span ?ts.
```

durch

```
?int rdf:type duomo:Record_Creation;
 ecrm:P4_has_time-span ?ts.
```

9.3.3 Weitere Analyseschritte

Im Folgenden gehen wir der Frage nach, wie die vorhandenen Daten weiter angereichert werden können. Unser Hauptaugenmerk liegt hier auf der Einbeziehung von extern zugänglichen Ressourcen, die als *Linked Open Data* zur Verfügung stehen. Ziel ist es, Fragen der folgenden Art beantworten zu können:

- Welche Materialien wurden wofür gebraucht?
- Welche Handlungen fanden in welcher Reihenfolge statt?

Materialien werden in der Datenbank in verschiedenen Kontexten benutzt. Wollen wir nach allen Vorkommnissen von Materialien in unterschiedlichen Formen suchen, müssen die im Text bzw. in der Datenbank vorkommenden Begriffe mit standardisierten und abstrakteren Konzepten verbunden werden. Dieses geschieht beispielhaft für zwei benutzte Materialien und Handlungen. Hierbei geht es

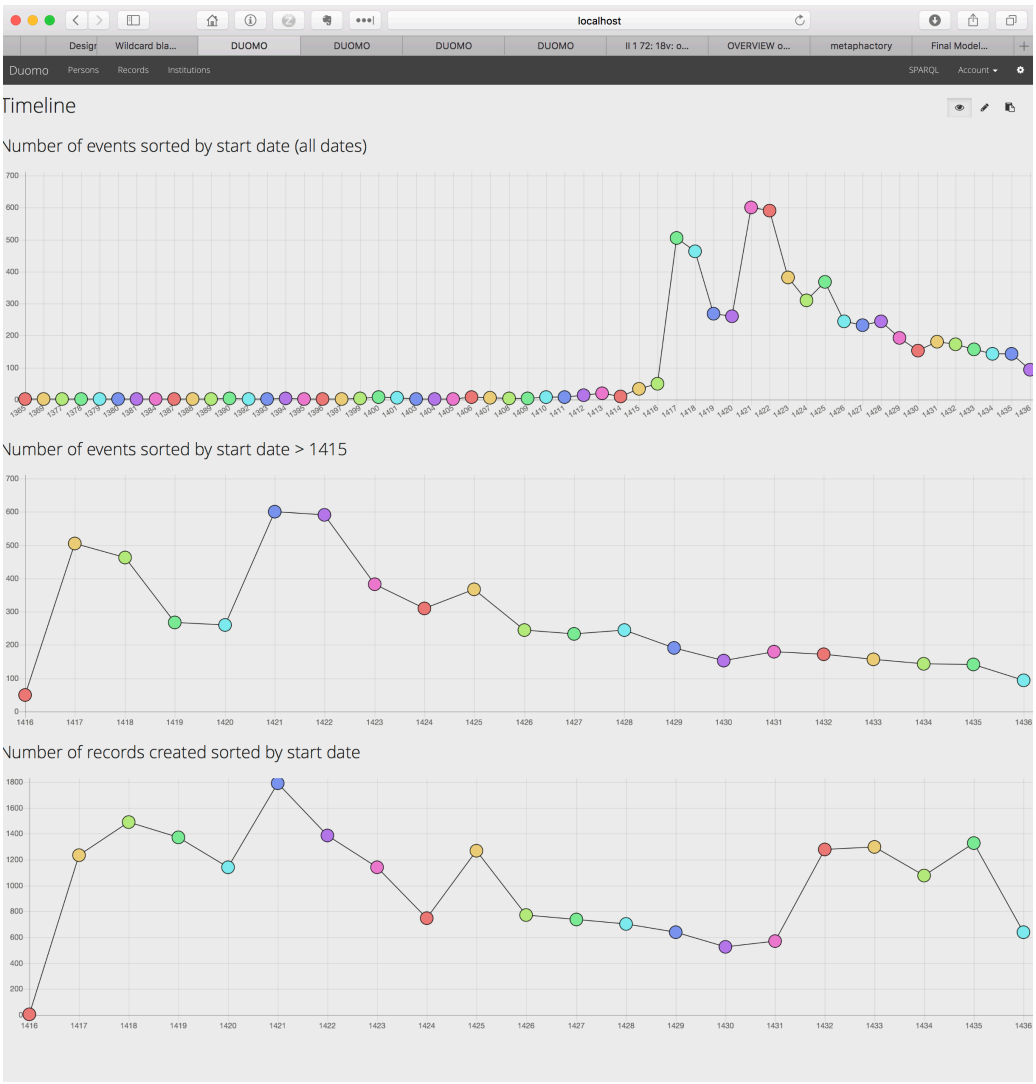


Abbildung 9.23: Visualisierung Anzahl von Duomo Events pro Jahr (oben), Anzahl von Duomo Events pro Jahr ab 1416 zum besseren Vergleich (mittig) und Anzahl der erzeugten Records pro Jahr (unten)

uns zu diesem Zeitpunkt um die Darstellung der Methode, weniger um die Exaktheit der Aussagen. Der Mechanismus zur Identifikation von Konzepten ist zurzeit lediglich rudimentär und macht nur wenige Annahmen über den Kontext von Begriffen, so dass Fehlzuordnungen zunächst noch häufig zu erwarten sind.[17] Zunächst gehen wir wie folgt vor:

- Einträge aus dem Thesaurus der analytischen Kategorien aus Abschnitt 9.2.4
- Begriffe aus den Regesten
- Auszeichnung von Kategorien in den Transkriptionen

9.3.4 Kontextualisierung der analytischen Kategorien und Regesten durch Verbindung zu WordNet

Sowohl die analytischen Kategorien als auch die Regesten liegen in englischen Übersetzungen vor. Das macht es in den ersten Schritten möglich, für die Kontextualisierung auf eine Reihen von offenen Hilfsmitteln und Diensten zurückzugreifen. Insbesondere können wir Begriffe mit Synonymgruppen und lexikalischen Einheiten in *WordNet* verbinden [153, 176], für die wiederum Verbindungen zu enzyklopädischen Ressourcen hergestellt werden können. Mittels des Paketes für die Analysen natürlicher Sprachen NLTK [158, 27] von Python werden dazu die Mengen synonymer Worte (Synsets) auf der Grundlage von WordNet 2.0 ermittelt und mit den entsprechenden Termini in WordNet verbunden. [18] Für die Formulierung der Verbindung der bisher erzeugten Ressourcen benutzen wir eine einfache Ontologie. Alle Klassen, die sich aus der Analyse der Originaldaten ergeben, sind Unterklassen von **duomo:Analytical_Entity** einer Unterklasse von **E28_Conceptual_Object**. Die oben eingeführten Ressourcen, die sich auf Wordnet beziehen, sind von der Klasse **duomo:Synset** und **duomo:Lexical_Entry**. Sie werden mit Ressourcen aus der Duomo-Datenbank über

- **duomo:is related to synset** und
- **duomo:is_primary_related_to_synset**, bzw.
- **duomo:is_related_to_lexicalEntry** und
- **duomo:is_primary_related_to_lexicalEntry**

[17] An dieser Stelle geht es in erster Linie um ein "Proof-Of-Concept"´. Zum Problem der ontologischen Annotationen siehe z.B.[212]. Mit Machine-Learning-Methoden und halbautomatischen Korrekturen könnte das Ergebnis sicher erheblich verbessert werden.

[18] Alle Synsetausdrücke lassen sich mit URIs für Ressourcen auf der Grundlage von WordNet 2 verbinden. Der Namespace für WordNet 2.0 ist hierbei `http://www.w3.org/2006/03/wn/wn20/instances/`. Ressourcen haben URLs der Form `http://www.w3.org/2006/03/wn/wn20/instances/synset-chair-noun-1` und werden über diese URL entweder in HTML-Darstellung [274] oder als RDF aufgelöst. Zur Zeit läßt sich WordNet zusätzlich über mehrere SPARQL-Endpunkte abfragen z.B. `http://wordnet.rkbexplorer.com/sparql/` [113] und `http://factforge.net/sparql` [77]. Aus Performanzgründen greife ich auf eine eigene Version von WordNet 3.0 von [177] zurück.

Neben Performanz ist auch die Frage nach der Stabilität von Links relevant. In der ersten Fassung dieser Arbeit hatte ich anstelle von `http://www.w3.org/2006/03/wn/wn20/instances/` noch `http://purl.org/vocabularies/princeton/` . . benutzt. Diese ergeben jedoch jetzt am 11. August 2017 einen 404 Fehler – werden also nicht mehr aufgelöst. Gleiches gilt zur Zeit (11. August 2017) für den Namespace für WordNet 3.1. Hier ist `http://wordnet-rdf.princeton.edu/ontology` nicht erreichbar.

(Abbildung 9.24) verbunden. Die Relation **duomo:is_primary_related_to_synset** zeichnet hierbei den primären Synonymsatz in WordNet aus. Mittels *analyseGraphs.analyse* erzeugen wir die entsprechenden Relationen und legen sie ab.[19]

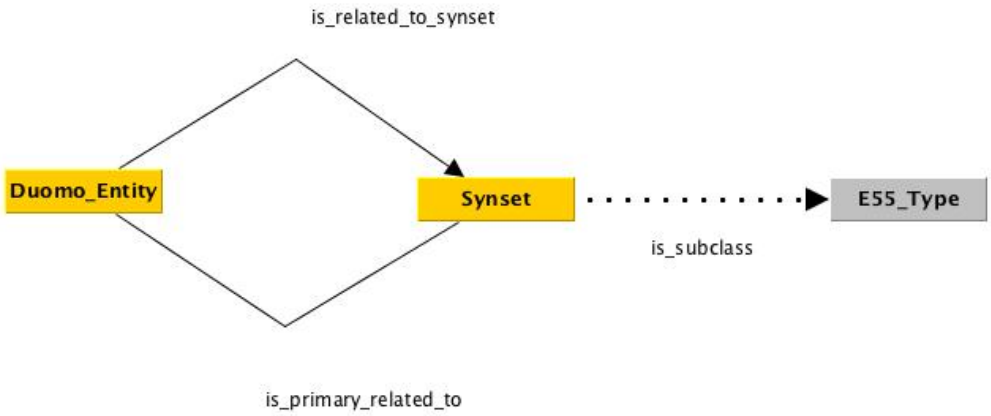


Abbildung 9.24: Synsets und Duomo Entities

Bei den Inhaltsangaben interessieren wir uns zunächst nur für Ausdrücke, in denen Handlungen näher beschrieben werden. Wir durchsuchen daher die Inhaltsangaben zunächst nur nach präpositionalen Ausdrücken. Dazu wenden wir eine einfache Grammatik auf der Grundlage regulärer Ausdrücke[20] auf alle Regesten an und erzeugen dadurch einen einfachen Aussagenbaum.

```
PP:   {<DT>?<JJ>?<VBN>?<PRP\$>?<NN|NNP|NNS|JJ|VB><NNS>?}
AP:  { <PP><IN><PP>}
APP: { <AP><IN|TO><AP|PP>}
C2C: {<APP|AP|PP>(<CC><PP>)?}
A2P: { <APP|C2C><IN|TO><C2C>}
```

Ein spezifisches Konstrukt behandeln wir hierbei besonders. Ausdrücke, in denen Funktionsbezeichnungen oder Namen mit Orten bzw. anderen Namen näher spezifiziert werden, wie *Commune of Pisa* oder *captain of Pisa*, werden zu einem Ausdruck zusammengefasst. Diese Ausdrücke filtern wir mit einer Grammatik und binden sie dann zu einem Term mittels „_“ zusammen. So erhalten wir z.B. für den Ausdruck:[21]

> Oath of warden and term of payment to debtor for the Commune of Pontedera for wine and butchering gabelles.

einen Baum der Form:

[19] *duomo:analyticalclassesSynSets*

[20] Python NLTK bietet dazu einen RegexParser an [169].

[21] `http://entities.mpiwg-berlin.mpg.de/duomo/2012071b-de66-4792-9ce5-791fd60f7541`

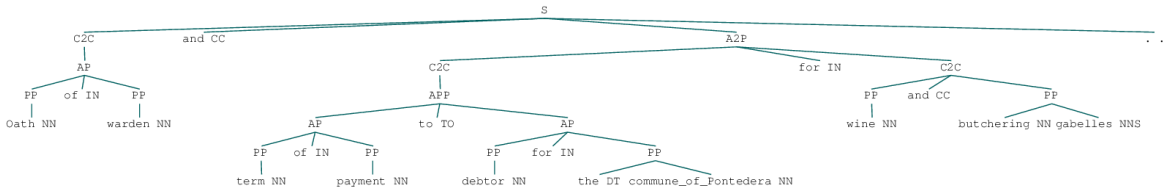


Abbildung 9.25: Baumdarstellung mit Hilfe einer einfachen Grammatik

Diese Bäume werden in eine einfache Ontologie (Abb. 9.26) übersetzt und das Ergebnis wird wieder er im Triplestore abgelegt.[22] Die genauere Darstellung findet sich wieder im Anhang.[23] Abbildung

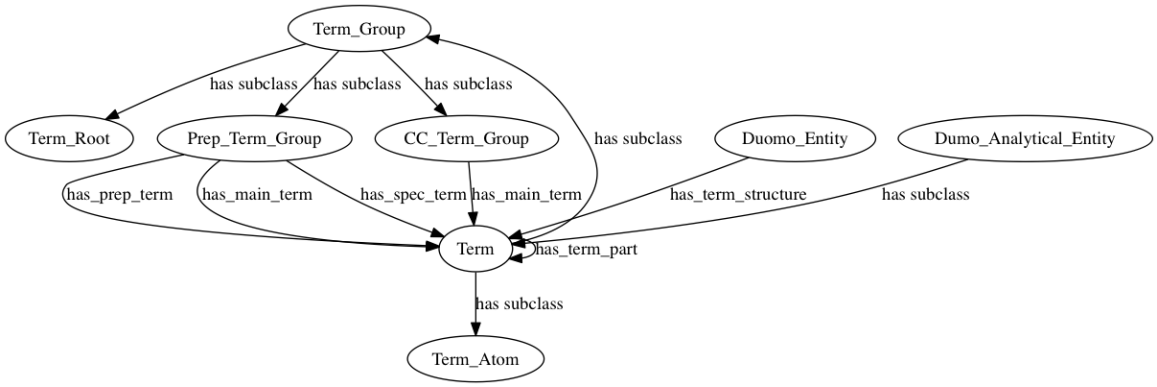


Abbildung 9.26: Einfaches Model für die Anbindung von Termen

9.27 zeigt den entsprechenden Teilgraphen für den Term von oben.

9.3.5 Struktur von WordNet in RDF

Wir benutzen WordNet 3.1 sowie 3.0 [153, 229] in der Version von [177], um die Einträge zu kontextualisieren. Da wir später möglichst viele Verbindungen auffinden wollen, verbinden wir alle Termini, die von NLTK als Synset ermittelt wurden, mit WordNet auf zwei unterschiedliche Weisen. Primär suchen wir alle kanonischen Formen **lemon:CanonicalForm**[24] für das Hauptwort. Zu diesen kanonischen Formen gehören jeweils lexikalische Einträge **lemon:LexicalEntry**. Diese verbinden wir mit den entsprechenden Ressourcen in unserem Triplestore (Abbildung 9.28). Außerdem verbinden wir die Ausdrücke direkt mit Synsets über die Label des Synsets (Abbildung 9.29). WordNet stellt zwischen lexikalischen Ausdrücken über Synonyme, Hypernyme und Hyponyme Beziehungen zwischen Begriffen her. Damit können komplexere Suchen realisiert werden, die über die eigentlich in der Datenbank vorhandenen Beziehungen hinausgehen.

Die Verbindung der Einträge in der Duomo-Datenbank mit den entsprechenden lexikalischen Ausdrücken erlaubt dann beispielsweise direkt Abfragen über miteinander in Beziehung stehende Konzepte. Abbildung 9.30 zeigt die Verbindung zwischen den lexikalischen Einträgen *metal* und *iron* bzw.

[22] *duomo:graphs/regesTermsAndSynSets*
[23] Siehe [267].
[24] Zum Lemon-Namespace siehe [141].

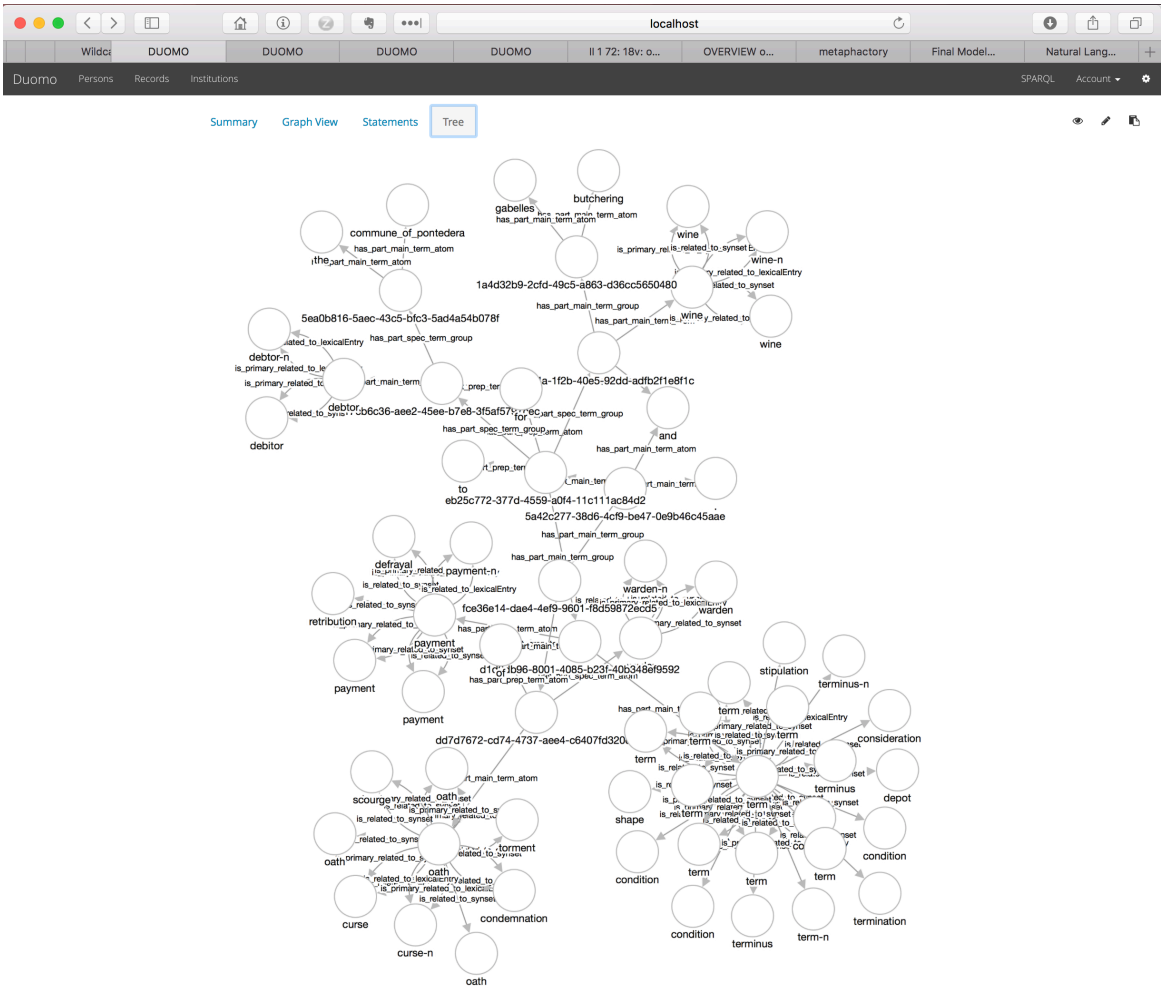


Abbildung 9.27: Beispiel für die Baumdarstellung einer der Regesten

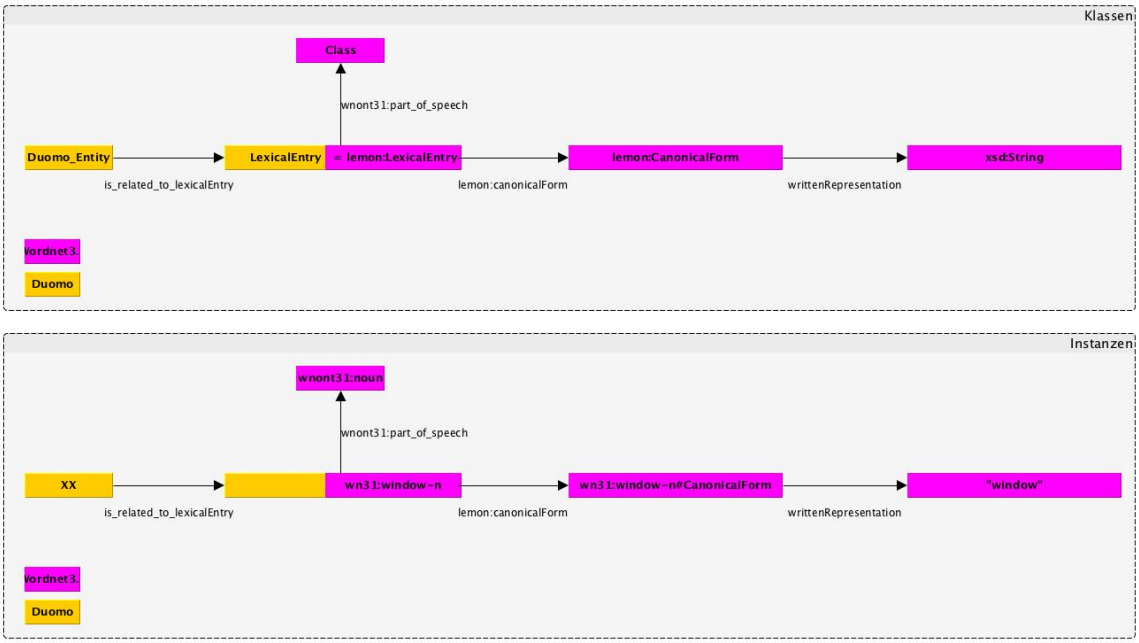


Abbildung 9.28: Verbindung von Duomo und WordNet über lexikalische Einträge

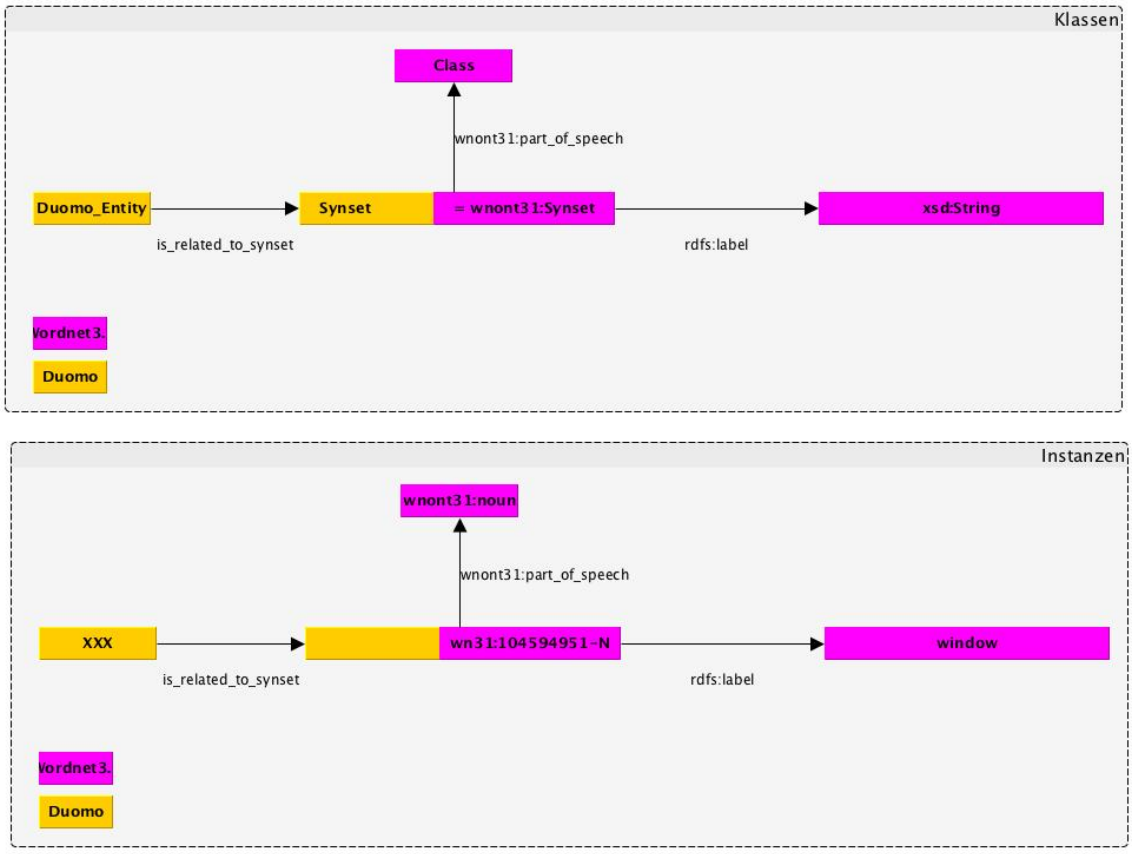


Abbildung 9.29: Verbindung von Duomo und WordNet über Synsets

brass. Messing (*brass*) als Legierung ist hier über einen anderen Synset von Metall – nämlich den eher alltäglichen Gebrauch auch für Legierungen mit dem Eintrag *metal* – verbunden, während *iron* ein direktes Hypernym von *metal* im Sinne der chemisch-physikalischen Definition von Metall ist. Eine

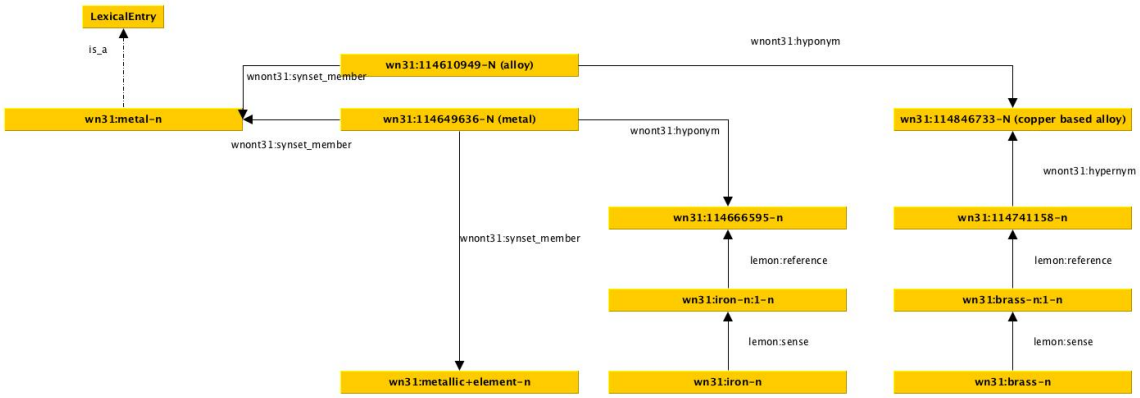


Abbildung 9.30: *metal* and *iron*

Abfrage über allen Vorkommnissen von Metallen lässt sich damit in SPARQL in der in Abbildung 9.31 dargestellten Form ausdrücken. Ein anderer Mehrwert ergibt sich durch die in WordNet direkt vorhandenen Übersetzungen einzelner Termini, so dass damit ein erster Schritt zu einer multilingualen Suche gegangen wird. Wir ergänzen die Daten um weitere Sprachen mittels der in [76] bereitgestellten Ressource und beziehen diese mittels **duomo:guess_lemma** auf die entsprechenden Synsets.[25]

An der TU Darmstadt [275] werden darüber hinaus Identifikationen zwischen WordNet und der Wikipedia zur Verfügung gestellt. Diese werden ebenfalls mit den entsprechenden Synsets über **duomo:guess_wikipedia** verbunden. [26]

9.3.6 Motivation – Linked Open Data, WordNet und Wikipedia

Eine zusätzliche Motivation für die Anbindung von wissenschaftlichen Datenbanken ergab sich aus der folgenden Überlegung. In [150] und [58] zeigen de Melo und Weikum, dass es mit vertretbarem Aufwand möglich ist, die Qualität von Links in *Wikipedia*, sowie die Verbindung von Wikipedia und WordNet mit Hilfe von Algorithmen auf den entsprechenden Wissensgraphen einzuschätzen und zu verbessern. Verbinden wir nun wissenschaftliche Datenbanken, denen wir zunächst einmal eine höhere Verlässlichkeit und damit Gewichtung zuweisen wollen etwa mit den Daten der Wikipedia, dann ließe sich auf diese Art und Weise die Qualität der Links dort weiter erhöhen. Umgekehrt ließe sich die Verbindung von Konzepten und Begriffen innerhalb der wissenschaftlichen Datenbanken gegenprüfen und darüberhinaus fehlende Beziehungen analog zu den Algorithmen für die Ergänzung der internen Wikipedia-Referenzen ergänzen. Es ergäbe sich also eine *win-win*-Situation für alle Beteiligten.

[25] Diese liegen jeweils in Graphen: *duomo:graphs/guess_lemma/TYPE/LANG*, TYPE = c ist hierbei der Datensatz gemäß [31] und TYPE = w derjenige aus [32], der italienische Datensatz innerhalb dieser Kompilationen stammt aus [241].

[26] Zur Erzeugung der Referenzen wurden zwei unterschiedliche Algorithmen benutzt und die entsprechenden Datensätze zur Verfügung gestellt. Wir binden in *duomo:graphs/guess_wikipedia/1* denjenigen aus [148] und *duomo:graphs/guess_wikipedia/2* den aus [273] ein.

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX : <http://ontologies.mpiwg-berlin.mpg.de/duomo/>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix duomo: <http://ontologies.mpiwg-berlin.mpg.de/duomo/>
prefix wn20: <http://www.w3.org/2006/03/wn/wn20/instances/>
prefix wn31: <http://wordnet-rdf.princeton.edu/wn31/>
prefix wn30: <http://wordnet-rdf.princeton.edu/wn31/>
prefix wnont31: <http://wordnet-rdf.princeton.edu/ontology#>
prefix lemon: <http://lemon-model.net/lemon#>
select distinct ?le ?y where {
            ?cf lemon:writtenRep "metal"@eng.
            ?s lemon:canonicalForm ?cf.

            ?synset wnont31:synset_member ?s;
                    wnont31:part_of_speech wnont31:noun;
                    wnont31:hyponym+ ?refs.
            ?sense  lemon:reference ?refs.
            {
   ?le lemon:sense ?sense.
}
   union
            {
             ?synset wnont31:synset_member ?le.
            }

    }
```

Abbildung 9.31: SPARQL Abfrage für alle Hypernyme zum Nomen *metal*.

```
select distinct ?le ?y where {
            ?cf lemon:writtenRep "metal"@eng.
            ?s lemon:canonicalForm ?cf.

            ?synset wnont31:synset member ?s;
                    wnont31:part_of_speech wnont31:noun;
                    wnont31:hyponym+ ?refs.
            ?sense  lemon:reference ?refs.
            {
   ?le lemon:sense ?sense.
}
   union
            {
             ?synset wnont31:synset_member ?le.
            }
          ?de
          duomo:is_related_to_lexicalEntry|duomo:is_primary_related_to_lexicalEntry ?le.
         ?a ?p ?del.
         select ?p where {?p rdfs:subPropertyOf duomo:has_term_atom.}
   ?pt duomo:has_part_term ?a.
    }
```

Abbildung 9.32: SPARQL-Abfrage für alle Hypernyme zu Nomen *metal*

Leider sind jedoch wissenschaftliche Datenbanken, die konsequent Anbindungen an entsprechende Ontologien vornehmen, noch selten. Ein Problem hierbei ist sicher auch die Zurückhaltung von Wissenschaftlern, sich auf die in offenen Datenbanken vorhandenen Konzepte zu beziehen, da hier eine doppelte Fehlerquelle gesehen wird: Einerseits ist die Übereinstimmung zwischen Konzepten aus externen Ontologien, Thesauri und Vokabularen mit den eigenen Konzepten immer nur in gewissen Grenzen möglich, andererseits wird offenen Ontologien nur in geringem Maße vertraut. Die oben genannten Arbeiten zeigen jedoch, dass trotz solcher Fehler ein Mehrwert erzeugt werden kann. Jedoch gilt auch hier: Dieser kann nur erreicht werden, wenn die Daten so offen wie möglich sind, so dass auf sie im Kontext von Forschungen in der Informatik einfach zurückgegriffen werden kann. Die Bereitstellung von Thesauri, wie sie mittlerweile das Getty Institut mit Open-Source-Lizenzen unternimmt, ist in dieser Hinsicht ein großer Schritt nach vorne [88].

9.3.7 Einfache Abfragen

Aus der Abfrage in Abbildung 9.31 lässt sich nun leicht die Abfrage nach allen Titelangaben, in denen Metalle vorkommen, konstruieren. Hierbei greife ich aus Performanzgründen auf das Subselect anstelle eines Tripels zurück. Außerdem wurde die Inferenz für **duomo:has_part_term** bezüglich dessen Unterrelationen explizit ausgeführt und im Triplestore gesichert. Die obige Abfrage erweitern wir zusätzlich noch um die Abfragen für die analytischen Kategorien, in denen Metall vorkommt, sowie um das Erstellungsdatums des Eintrages. So bekommen wir nun wieder direkt mit Hilfe der Visuali-

```
SELECT ?year (COUNT(distinct ?record) as ?record_n)
where {
{
select distinct ?record ?year  where {
            ?cf lemon:writtenRep "metal"@eng.
            ?s lemon:canonicalForm ?cf.

            ?synset wnont31:synset_member ?s;
                    wnont31:part_of_speech wnont31:noun;
                    wnont31:hyponym+ ?refs.
            ?sense  lemon:reference ?refs.
            {
   ?le lemon:sense ?sense.
}
   union
            {
             ?synset wnont31:synset_member ?le.
            }
          ?de
       duomo:is_related_to_lexicalEntry|duomo:is_primary_related_to_lexicalEntry ?le.
            {
              ?a ?p ?de.
              {
              select ?p where {?p rdfs:subPropertyOf duomo:has_term_atom.}
              }
              ?pt duomo:has_part_term* ?a.
              ?pt  a duomo:Term_Root.
              ?reges duomo:has_term_structure ?pt.
              ?record duomo:has_summary ?reges.

       } union {
         ?st duomo:has_analytical_type ?de.
         ?event duomo:expresses_statement ?st.
         ?record duomo:has_interpretation ?event.
       }
       ?record duomo:has_date_of_writing ?date.
   ?date duomo:has_XSD_date ?xsddate.
            bind( year(?xsddate) as ?year)
    filter (?year < 1900)

    }
}
}
group by ?year
order by ?year
.
```

Abbildung 9.33: Abfrage nach Metallen in Titeln und analytischen Kategorien gruppiert nach Verfassungsjahren der Einträge. Das Subselect-Strukturen wurden hier eingeführt, da *blazegraph* keine Wildcards in RDF-Pfaden zuläßt, wenn diese gruppiert werden sollen.

sierungsmöglichkeiten von *metaphactory* die zeitliche Verteilung der Einträge (Abbildung 9.33).

Für die Abfrage von Events über Zeiträume ergeben sich wieder erhebliche Performanzunterschiede je nach Aufteilung und Reihenfolge der Tripel in der SPARQL-Abfrage. Eine Abfrage ohne Subselects hat in meiner Konfiguration nach einer Stunde zu keinen Ergebnissen geführt, während die Abfrage mit den Subselects wie in Abbildung 9.33 nach fünf Sekunden zu Ergebnissen führte. Die Visualisierung dazu findet sich in Abbildung 9.34.

9.3.8 Personennetzwerk in den Daten

Wir benutzen auch in diesem Falle unsere Daten, um Beziehungen zwischen Personen zu ermitteln. Im einfacheren Fall sind Personen sind immer dann miteinander verbunden, wenn sie in einem Eintrag gemeinsam erwähnt werden. Aus unseren Daten können wir dieses Netzwerk wieder generieren. Aus

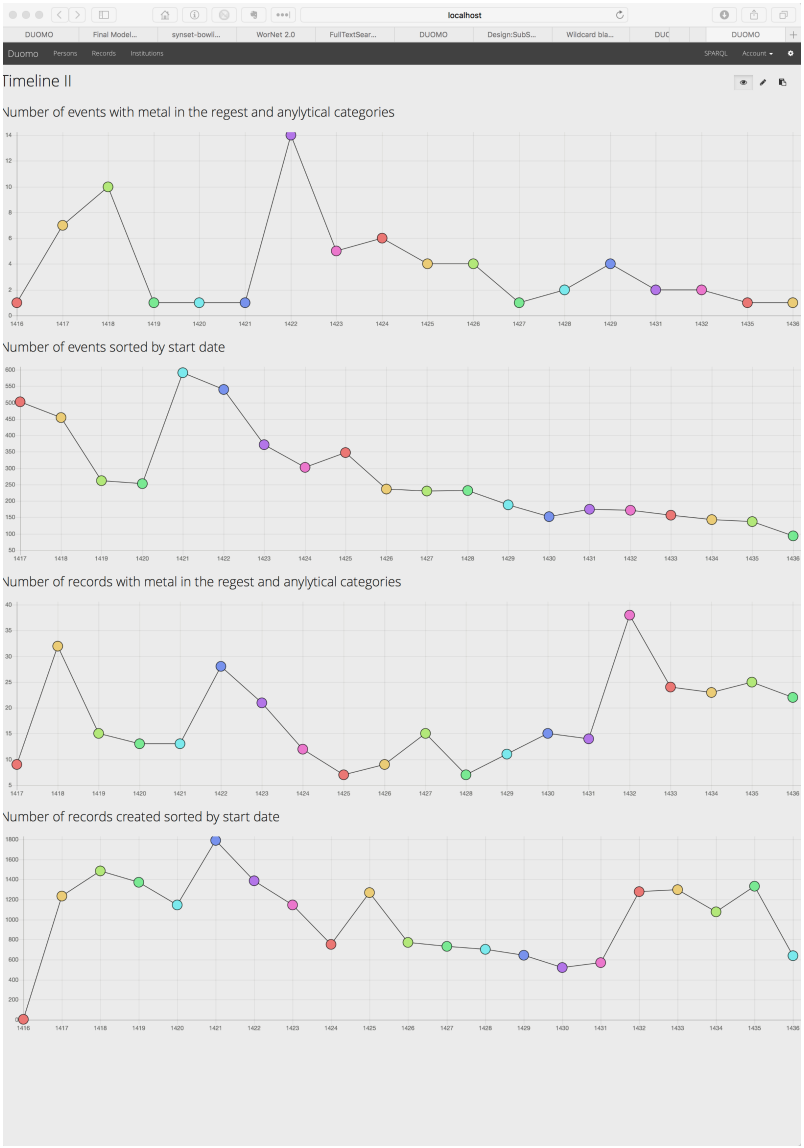


Abbildung 9.34: Zeitlicher Verlauf der Ereignisse bezogen auf *metal*

unserer Ontologie ergeben sich Personenbeziehungen bezüglich eines Ereignisses wieder unmittelbar (Abbildung 9.35). Mit einer entsprechenden SPARQL-Abfrage (Abbildung 9.36) erzeugen wir das

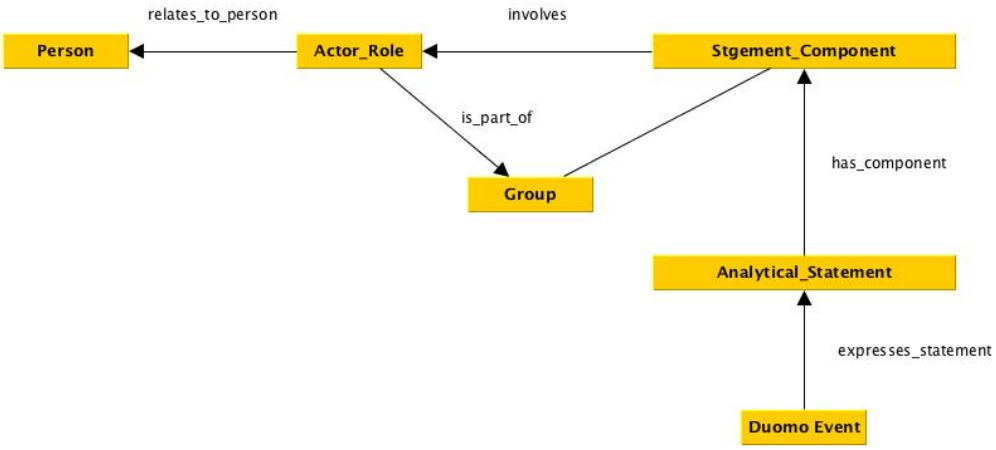


Abbildung 9.35: Personen und Events

entsprechende Netzwerk. Auch hier können wir die Zeiten hinzufügen und eine dynamische Entwicklung des Netzwerkes über die Jahre darstellen. Das bipartite Netzwerk (Abb. 9.37) aller Einträge und Personen besteht dann im Wesentlichen aus einer großen Komponente mit 10329 von 11835 Knoten, wobei wir hier nur die Ereignisse einbeziehen, die Personen miteinander verbinden – also solche mit einem Degree größer als Eins. Insgesamt hat der Graph bisher 27693 Knoten. Wir können auch hier wieder Daten im Jahresgraphformat erzeugen und entsprechend visualisieren. Das Beispiel in Abb. 9.38 zeigt die Entwicklung des Personennetzwerks in drei Jahresabständen und jeweils für Zeiträume von drei Jahren und jeweils die größte Komponente.

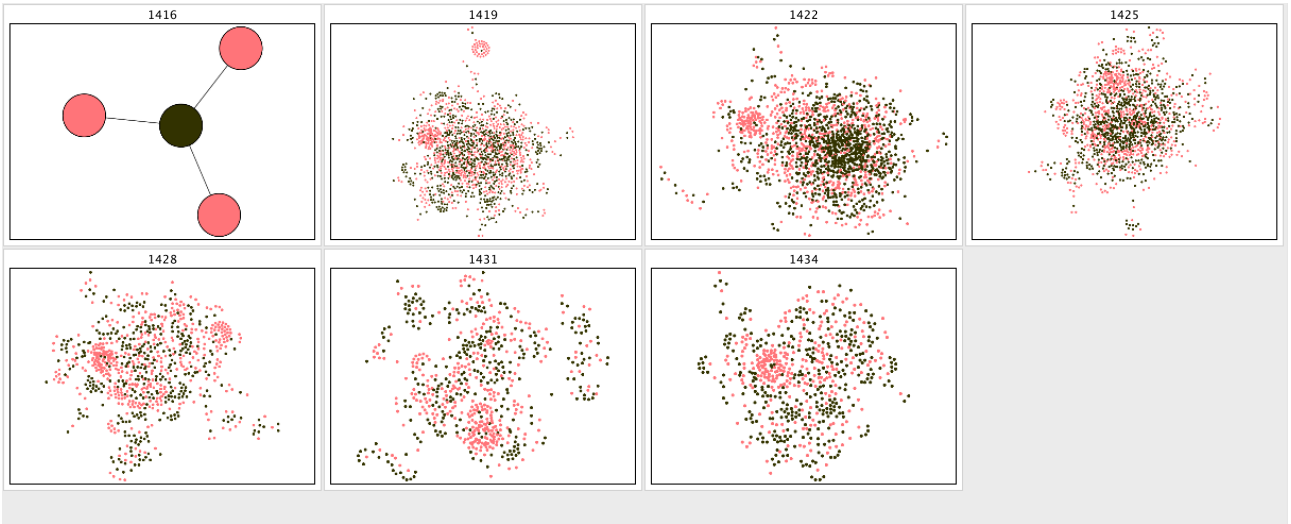


Abbildung 9.38: Bipartiter Graph der Einträge und Personen in der Datenbank für jeweils 3 Jahre, Ereignisse sind hier schwarz dargestellt.

```
PREFIX efrbroo: <http://erlangen-crm.org/efrbroo/>
PREFIX duomo: <http://ontologies.mpiwg-berlin.mpg.de/duomo/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX crm: <http://erlangen-crm.org/160714/>

SELECT distinct ?ns ?nm ?ns__en ?ns__date ?ns__xsdate WHERE {
 ?rec crm:P48_has_preferred_identifier/rdfs:label ?ns;
                duomo:has_summary/rdfs:label ?ns__en.
         ?ce efrbroo:R17_created ?rec.
         ?ce crm:P4_has_time-span ?ts.
          ?ts rdfs:label ?ns__date.
      optional{
    ?ts duomo:has_XSD_date ?ns__xsdate.}

 ?rec duomo:has_interpretation ?interp.
{ {
 ?interp duomo:expresses_statement ?stm.
 ?stm duomo:has_component ?cmp.
 ?cmp duomo:involves ?ar.
 ?p duomo:is_part_of_group* ?ar.
} union  {
 ?p duomo:is_mentioned_in ?interp.
}}
 ?p duomo:relates_to_person ?y.

 ?y rdfs:label ?nm.
}
LIMIT 1000000
.
```

Abbildung 9.36: Erstellung des Personennetzwerkes

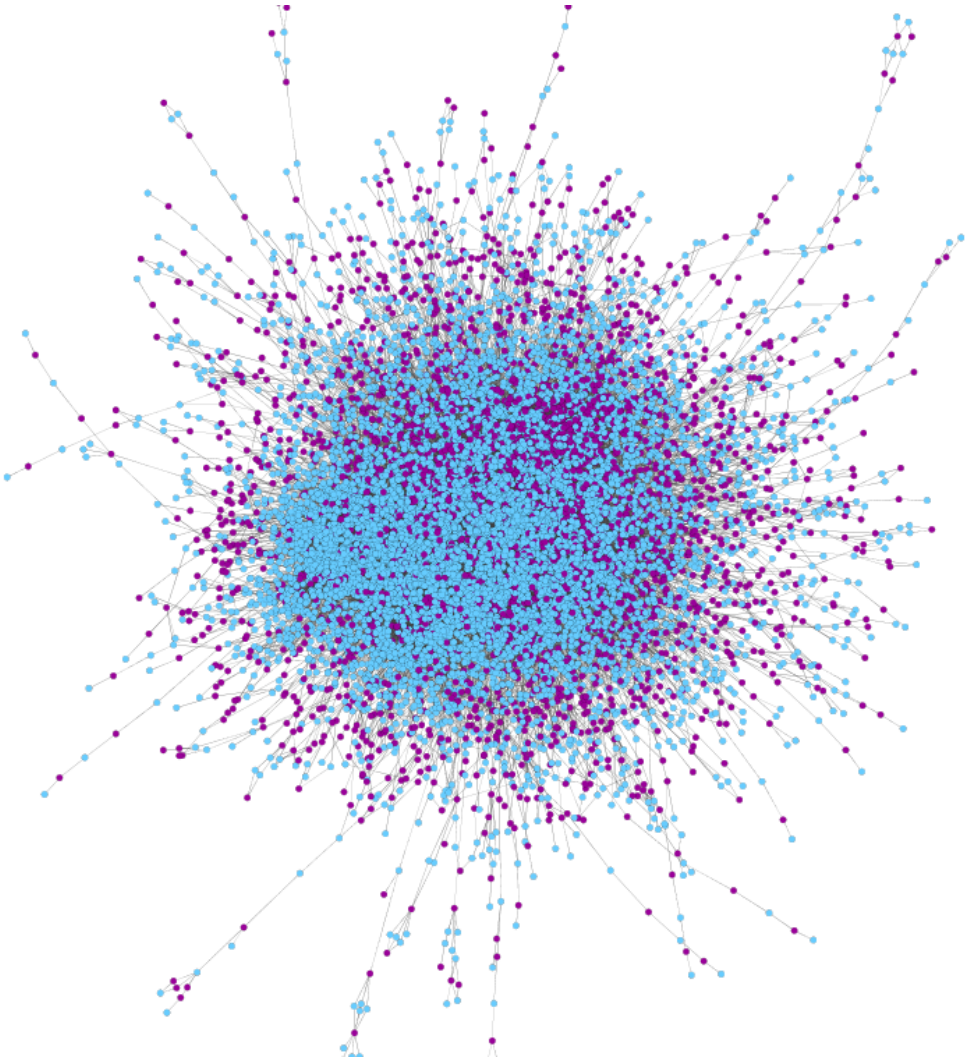


Abbildung 9.37: Bipartiter Graph der Einträge und Personen in der Datenbank, Ereignisse sind hier blau.

9.4 Zusammenfassung und Ausblick

In diesem Kapitel kann aus Platzgründen nur ein kleiner Teil der Möglichkeiten, die sich durch die Transformation der vorhandenen Daten in eine ontologiebasierte Lösung eröffnen, vorgestellt werden. So lassen sich nun leicht Abfragen über Kontexte erstellen. Einige dieser Schritte wären auch in einer relationalen Datenbank möglich, wie etwa die Reihenfolge von Ereignissen. Grundsätzlich einfacher sind jedoch Abfragen, die darauf abzielen, den Zusammenhang zwischen Ereignissen aufzufinden. Wir können zum Beispiel Personen ausmachen, die Ereignisse miteinander verbinden, wie wir in Abschnitt 9.3.8 in der einfachen Form eines Personennetzwerkes zeigen. Multilinguale Zugänge und Zugänge über Konzepte lassen sich durch das Anreichern der Daten mittels WordNet und Wikidata erreichen. In 9.3.4 haben wir dies skizziert. In der geschilderten einfachen Form der Identifikation ergeben sich viele Ambiguitäten, die noch aufgelöst werden müssen. Teilweise kann dies durch die Einbeziehung der unmittelbaren Kontexte der Einträge erfolgen, teilweise wird dieses halbautomatisch erfolgen müssen. Da es den Rahmen dieser Arbeit sprengt, sind die ersten Schritte, die wir in diese Richtung gegangen sind, hier nicht dargestellt. So eröffnet sich durch Topic-Modelling in Kombination mit der manuellen Auszeichnung, die bereits in Form der Indizes vorgenommen ist, die Möglichkeit die Synonymgruppen weiter einzugrenzen. Nicht dargestellt ist auch der noch immer nicht erfolgreiche Versuch, die extrahierten Daten mit möglichen anderen Projekten, die Daten als Linked Open Data zur Verfügung stellen, zu verbinden. In Arbeit ist zur Zeit eine Verbindung der Daten mit *Art and Architecture Thesaurus Online (AAT)* [13, 88]. Dies würde es dann ermöglichen, neben Oberbegriffen, wie sie WordNet, erlaubt auch nach Handlungsebenen zu suchen, die im AAT aus kunsthistorischer Sicht klassifiziert und mit entsprechenden Begriffen verbunden sind. Eine Frage, die damit beantwortet werden könnte, wäre dann die Frage nach typischen Handlungsreihenfolgen, wie Planung – Anschaffung – Nutzung, ihren Zeitskalen und Abhängigkeiten.

Kapitel 10

Die Sphaera des Sacrobosco als epistemisches Netzwerk

Die folgende im Vergleich zu den anderen Fallstudien kurze Darstellung unserer ersten Arbeiten an der Sphaera des Sacrobosco schließt die Reihe der Beispiele, die die Arbeitsweise an historischen Projekten unter Einsatz von Methoden aus der Netzwerktheorie und der semantischen Modellierung deutlich machen soll, ab. Modellierung und Netzwerkanalyse am Beispiel der Sphaera war eine der ersten Arbeiten unter Einbeziehung sozialer Netzwerktheorie und weiterer Ideen aus der Netzwerkforschung. Inspiriert waren diese Arbeiten zu diesem Zeitpunkt vor allem durch Thomas W. Valentes Buch *Network Models of the Diffusion of Innovations* [246]. In [204] haben wir ausführlich erläutert, wie diese Fallstudie sich in den allgemeinen Kontext einordnet. Dieses soll im Folgenden nicht wiederholt werden. Obwohl dieses Projekt chronologisch eines der ersten Projekte war, das mit Hilfe des geschilderten Ansatzes angegangen wurde, ist es zugleich das Projekt, das die noch offenen Forschungsfragen deutlich zu Tage treten lässt. Insofern bietet es sich als Übergang zum letzten Kapitel an.

Die Darstellungen dieses Kapitels beziehen sich auf die erste Phase der gemeinsamen Arbeit mit Matteo Valleriani. Florian Kräutli und Matteo Valleriani haben mittlerweile die Teile des Modellierungsansatzes, der sich auf die bibliographischen Aspekte des Projektes bezieht, in [135] veröffentlicht, so dass dieses hier nicht wiedergeben werden muss. Eine ausführliche historische Einordnung der Netzwerke wurde von Matteo Valleriani in [247] veröffentlicht.

Eine Reihe der Überlegungen, von denen wir zu Beginn des Projektes ausgegangen sind, verfolgen wir zurzeit so nicht mehr. Für die Entwickler der Methodologie sind sie jedoch nach wie vor zentral.[1] Die Abbildungen und Auswertungen beziehen sich im Folgenden auf den Datenbestand von 2015. Da der geschilderte Prozess jedoch in weiten Teilen paradigmatisch für das Zusammengehen von Informatik und Geisteswissenschaften ist und der grundlegende Ansatz der Zusammenführung von Überlegungen im Rahmen der Netzwerktheorie und der Modellierungstheorie nach wie vor korrekt ist, soll dieses Beispiel vor allem verdeutlichen, wie der in Kapitel 5.8 beschriebene Ansatz einer

[1] Im Sinne von Karl Poppers Fußnote zu Abschnitt 77 in [186] gehe ich davon aus, dass gerade die Dokumentation von Irrtümern und Wirrungen wichtig für die Methodenentwicklung ist. Gerade die Digital Humanities leiden darunter, dass Ansätze, die nicht direkt zum Ziel geführt haben, oft unterschlagen werden, obwohl dieses zum Tagesgeschäft gehört.

Multilevelnetzwerktheorie zu einer systematischen Analyse historischer Prozesse führen kann. Sie ermöglicht es, historische Quellen besser einzuordnen, und macht zugleich Desiderata deutlich.

10.1 *Distant* und *close reading*

In den Digital Humanities hat sich das vor allem von Franco Moretti [157] in die Diskussion eingebrachte Konzept des *distant reading* mit seinem Counterpart dem *close reading* als Ansatz für die systematische Auswertung größerer Textkorpora etabliert. Auf die Details und die damit verbundene kritische Debatte können wir hier nicht eingehen.[2] Im Sinne des in dieser Arbeit immer wieder vertretenen Standpunktes einer pragmatischen Anwendung digitaler Methoden als heuristisches Hilfsmittel für die historische Forschung nehmen wir das Konzept hierbei insofern auf, dass neben die Detailstudie und das genaue Lesen der Quellen – also das *close reading* – eine Methodik gestellt wird, die eine große Anzahl von ähnlichen Quellen in strukturelle Beziehung zu den detailliert untersuchten Quellen stellt. Bei den in dieser Fallstudie benutzten Quellen – frühneuzeitliche Drucke – ist eine Volltexterfassung bisher durch Optical Character Recognition (OCR) nicht mit ausreichender Genauigkeit möglich, um sprachliche Veränderungen im Detail zu analysieren. Gesucht sind also Indikatoren, die die Texte auf Grund anderer Strukturmerkmale jenseits des exakten Textvergleiches in Beziehung zueinander setzen können und trotzdem eine inhaltliche Bewertung ermöglichen. Die im Folgenden nur kurz dargestellten Ansätze sind in diesem Sinne eine Annäherung an das *distant reading* im Falle von Textkorpora, die sich nicht oder nur schwer in Form maschinenlesbarer Volltexte erschließen lassen.[3]

10.2 Historischer Kontext

Der historische Kontext ist in [247, 57] dargestellt. Daher hier nur die für das Folgende wesentlichen Überlegungen. Die Sphaera des Sacrobosco ist ein erstmalig als Handschrift in der ersten Hälfte des 13. Jahrhunderts erschienenes grundlegendes Traktat, das in die Himmelsmechanik einführt und die dahinter liegenden mathematischen und naturphilosophischen Grundlagen erklärt. Es wurde für nahezu zwei Jahrhunderte als Unterrichtsmaterial an den sich in dieser Zeit etablierenden Lehreinrichtungen – vor allem den Universitäten – im gesamten europäischen Raum eingesetzt. Das Manuskript erschien vermutlich erstmalig in der zweiten Hälfte des 15. Jahrhunderts im Druck und wurde dann immer wieder an verschiedenen Stellen Europas neu aufgelegt. Im Verlauf dieser Neueditionen wurde der eigentliche Kerntext meist unverändert gelassen, jedoch wurden zusätzliche Beiträge weiterer Autoren mit in die unterschiedlichen Ausgaben aufgenommen. Diese Ergänzungen stehen in den weiteren Untersuchungen stellvertretend für neu aufgenommene Wissensbereiche. Diese dienen dann als Kennzeichen für den Wandel des mit der Sphaera verbundenen Wissenssystems dienen können. Welche Autoren zusätzlich aufgenommen wurden, hängt – so die Annahme – von der intendierten

[2] Zur kritischen Diskussion dazu siehe z.B. [90] und [209].

[3] Noch mehr gilt dieses für die gesamte Quellengattung handschriftlicher Notizen und von Manuskripten, ohne die historische Forschung nicht denkbar ist.

Leserschaft der jeweilig neuen Auflagen ab. Die Zusammensetzung der einzelnen Auflagen gibt daher einen Hinweis auf das dem Kreis der Herausgeber zugängliche Wissen. Neu in einer Ausgabe zusammengefasste Traktate stehen damit für einen neuen, vorher nicht explizit hergestellten Bezug zwischen unterschiedlichen Bereichen des wissenschaftlichen Wissens der Epoche. Die geographische Verbreitung ist ein Beispiel für Wissensdiffusion im gesamten europäischen Raum. Ermöglicht wurde dies unter anderem durch die lateinische Sprache als das universelle Medium der gelehrten Kommunikation im spätmittelalterlichen und frühneuzeitlichen Europa. Dadurch war ein Austausch von Wissen in schriftlicher Form zwischen unterschiedlichen Zentren in unterschiedlichen Ländern ohne sprachliche Barrieren möglich. Zugleich erschien das Werk in verschiedenen Übersetzungen in die Alltagssprachen der Zeit, wodurch eine Öffnung des Wissenssystems der Sphaera von akademischen Zirkeln hinein in breitere Schichten möglich wurde. Mit diesen Grundannahmen erhalten wir eine Struktur, die wir in das dieser Arbeit zugrundeliegenden Netzwerkmodell der Wissensentwicklung einordnen können. Die Identifikation der verschiedenen Ebenen wird im Folgenden geschildert.

10.3 Annahmen über die Struktur des Netzwerkes und dessen Modellierung

Im Sinne unseres in Kapitel 3.2 und 5.13 geschilderten Multilevelansatzes gilt es, Knoten auf den verschiedenen Ebenen des epistemischen Netzwerkes zu identifizieren. Nehmen wir jede einzelne Ausgabe der Sphaera als zeitlich-räumliche Manifestation des Wissenssystems der Sphaera, so wird die Binnenstruktur der jeweiligen Edition also die interne Anordnung der einzelnen Traktate zu einem Teilnetz des *semiotischen Netzwerks*, wenn wir die hinzugekommenen Traktate als Repräsentationen von neuen Wissensbereichen verstehen. Analog interpretieren wir eine mögliche Neuanordnung und Kürzungen bestehender Texte als Ausdruck einer veränderten Bedeutung bestehender Bausteine des Wissenssystems. Durch diese Interpretation verbinden wir die formale Struktur der Editionen mit dem semantischen Netzwerk, dem Wissenssystem der Sphaera. Explizit geschieht dies durch die Zuordnung von Texten und Textteilen zu Themen bzw. Themenfeldern, wie etwa der sphärischen Geometrie, der Nautik und astronomischer Theorien. In dieser Interpretation drückt jede neue Edition eine Mikrostruktur des Wissenssystems aus, lokalisiert in Raum und Zeit durch den Entstehungskontext der Edition, während die Verbindung der Editionen Aufschluss über die Makrostruktur des Wissenssystems geben kann.

10.3.1 Konkrete Umsetzung in eine Netzwerkstruktur und ein Modell

Abbildung 10.1 zeigt die Ausgangslage. Die Knoten des sozialen Netzwerkes ergeben sich aus einigen grundlegenden Annahmen des zeitgenössischen Lehrbetriebs und des bekannten Informationsaustausches. So besteht das Netzwerk aus den Autoren (Aut) der entsprechenden Teile (Part) des Werkes und den Verlegen (Verl) der einzelnen Ausgaben (Edit). Darüberhinaus sind die Abnehmer des Werkes, die Professoren (Prof) und Studenten(nicht in der Abbildung) an den einzelnen Orten (Plac), interessierte Laien (nicht in der Abbildung), sowie unterschiedliche Formen von Patronage

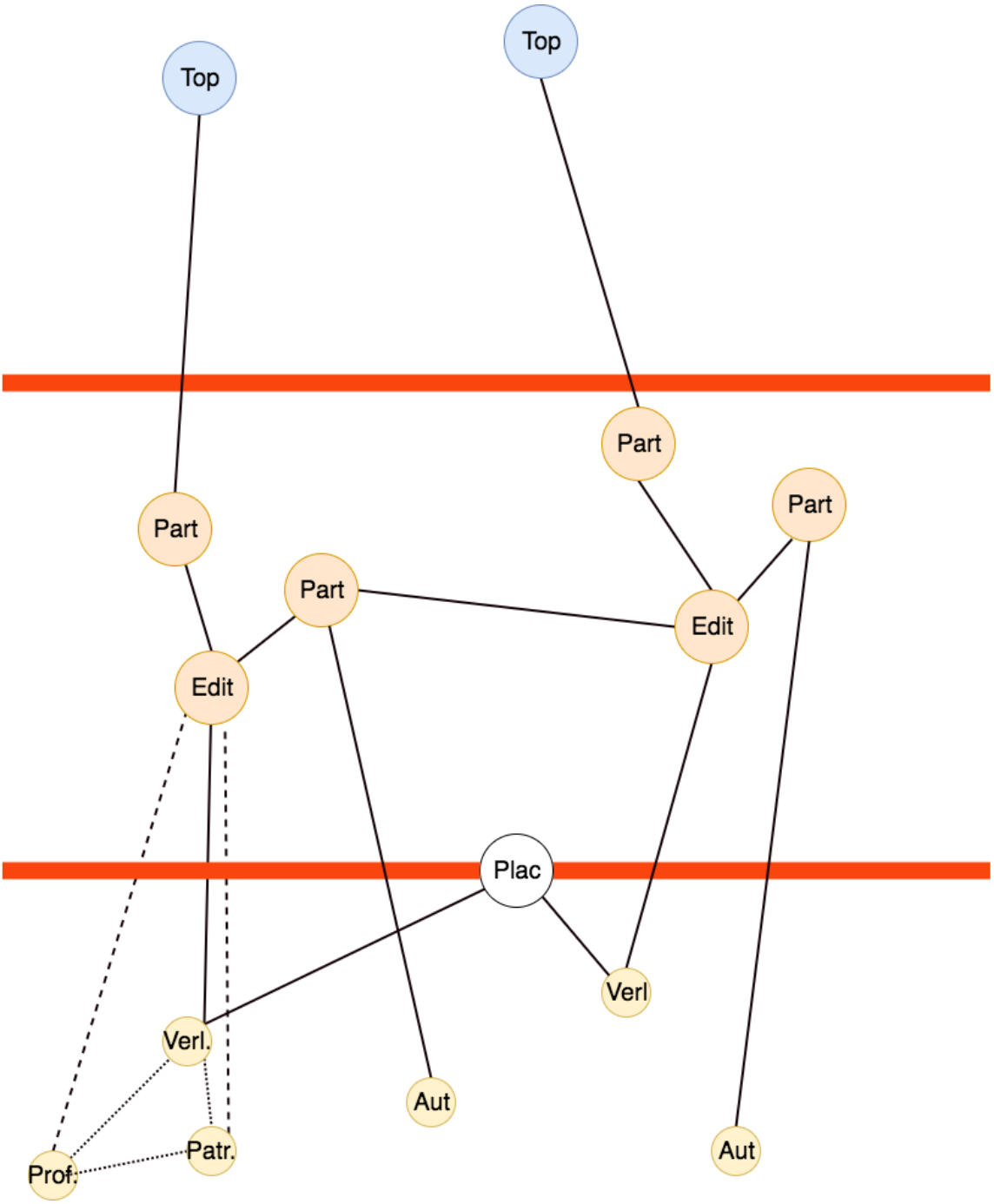


Abbildung 10.1: Struktur des epistemischen Netzwerkes der Sphaera

(Patr) Bestandteile des Netzwerkes. Die gestrichelten und gepunkteten Linien sind abgeleitete Verbindungen, die sich aus Annahmen über Beziehungen innerhalb des Netzwerkes ergeben. Auf diese werden wir in den nächsten Abschnitten weiter eingehen. Grundlegend für das Zustandekommen der einzelnen Ausgaben sind die Verleger bzw. Drucker an den verschiedenen Orten. Sie sind als aktiv Handelnde Teil des sozialen Netzwerkes und stehen in direkter Beziehung zu den Ausgaben, die durch ihr Handeln entstanden sind. Über denselben Verleger sind einzelne Ausgaben untereinander vernetzt. Über die Verbindung von Verlegern zu den Orten ihrer Tätigkeit ergibt sich ein erstes, stärker zusammenhängendes Netz. Wie oben geschildert, ist die Ausgangsannahme, dass die Ausgaben ein Ergebnis komplexer (bewusster oder unbewusster) Aushandlungsprozesse sind, an denen unterschiedliche Akteursgruppen beteiligt sind, wobei der Verleger dann die konkrete Aufgabe des Druckens übernimmt und zugleich, solange es keine reine Auftragsarbeit ist, auch das unternehmerische Risiko trägt. Daher liegt es nahe, das Modell zu erweitern und die einzelnen Editionen als Ergebnis einer Kette von Ereignissen darzustellen. Dazu führen wir als Konzept ein „Druckereignis“ als zentralen Ankerpunkt ein. In Abbildung 10.1 sind daher die Editionen in diesem Sinne zu verstehen. Dies eröffnet die Möglichkeit, weitere Aspekte des Zustandekommens der Edition schrittweise zu ergänzen, so dass dann letztendlich das Zustandekommen einer neuen Ausgabe das Ergebnis einer ganzen Reihe von Prozessen darstellt.

Die konkret mit dem Druck zusammenhängenden Ereignisse lassen sich in *FRBRoo*[4] ausdrücken, wie etwa die Konzeption, das Erstellen des Manuskriptes und der Druckvorlagen, sowie der erste Druck. Je mehr Details bekannt werden, desto komplexer wird die Struktur. Unser erstes Modell für die Beschreibung dieser Prozesse entstand zunächst zwar inspiriert von CRM, jedoch ohne eine explizite Anlehnung. Es zeigte sich jedoch schnell, dass *FRBRoo* weite Teile des Modell abdecken kann. Florian Kräutli hat das entsprechenden Modell für die Sphaera auf dieser Grundlage erstellt und dieses ist in [135] dokumentiert. Die einzelnen Ausgaben der Sphaera liegen mittlerweile in dieser Form beschrieben vor. Auf diese kann mittels *ResearchSpace*, einer Erweiterung von *metaphactory*[5], direkt zugegriffen werden.

Im Folgenden konzentrieren wir uns daher auf die für die Frage nach der mit der Sphaera verbundenen Transformation und die für die Ausbreitung des Wissenssystems relevanten Modellierungsschritte. Wie auch in den anderen Beispielen dieser Arbeit standen am Anfang einfache empirische Fragen im Vordergrund. Ziel war es, einen ersten zusammenhängenden Überblick über das Publikationsgeschehen zu bekommen und auszuloten, welche Ergebnisse sich mit vertretbarem Aufwand mittels des neuen Ansatzes erzielen lassen.

10.3.2 Einfache Netzwerke

Aus den vorliegenden Metadaten lassen sich auf der Grundlage der im Triplestore vorliegenden Daten Mutlilevel-Netzwerke erstellen, die einen ersten Eindruck vom epistemischen Netzwerk der Sphaera

[4] Siehe Abschnitt 4.4.
[5] Siehe Abschnitt 6.6.

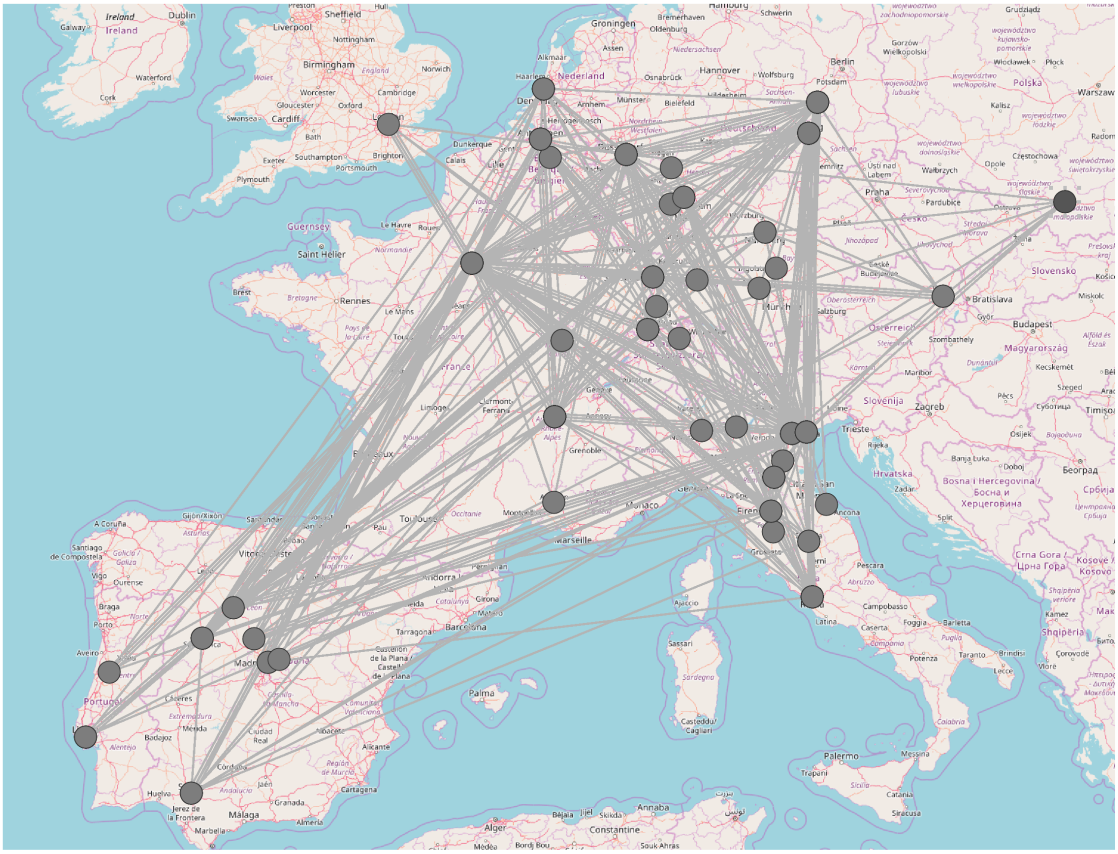


Abbildung 10.2: Gesamtübersicht nach Verlagsorten (Stand Dezember 2015, Visualiserung mit Visione, Karte von OpenStreetMap)

geben.[6] Wir beginnen mit einem einfachen Netzwerk, das sich aus den vorhandenen Metadaten über die einzelnen Editionen der Sphaera ergibt. Neben den Autoren- und Verlegerbeziehungen ergibt sich eine Halbordnung der einzelnen Ausgaben in Form einer **war_vor** Beziehung. Verbunden mit den geographischen Angaben des Erscheinungsortes ermöglicht dies einen ersten visuellen Eindruck einer möglichen Ausbreitung des Werkes und des damit verbundenen Wissenssystems. Auch hier gilt wieder, dass die Bearbeitung der Ausgangsdaten mittels eines Notizbuches geschieht.[7]

10.3.3 Ausbreitungsgeschichte der Sphaera

Die Analyse dieser Ausbreitung ist in [204, S. 51ff] beschrieben. Sie zeigt nach einer ersten Verzögerungsphase eine schnelle Verbreitung über Europa hin zu verschiedenen Zentren und schließlich gegen Ende der Verbreitungsgeschichte einen Rückgang auf nur noch einen Ort, den Stadtstaat Venedig. Diese Visualisierungen weisen zugleich auch auf die Gefahr einer solchen Darstellung hin. Ohne eine genaue Angabe, wie die Kanten zu lesen sind, birgt diese Darstellung die Gefahr einer Missdeutung. Beschrieben wird nicht, wie die Ausbreitung stattgefunden hat, sondern lediglich der

[6] Wie aus den im Triplestore vorliegenden semantischen Daten Netzwerke erzeugt werden können, wurde in den vorhergehenden Fallbeispielen ausgeführt.

[7] *sphaera_analysis_v0.8.ipynb*

Rahmen einer solchen Ausbreitung. Ein direkter Zusammenhang zwischen den Ausgaben wird durch diese Pfeile nicht unterstellt. Sie stellen lediglich die Randbedingungen für die Ausbreitung dar. Auch wenn Zeitordnung hier ein zugegeben einfaches Beispiel ist, ist diese jedoch eine grundlegende Voraussetzung für die Interpretation der Darstellung von Netzwerken, in der notwendige und hinreichende Bedingungen, Kausalität und Korrelation durch die Bildsprache häufig vermischt werden.

Es zeigt jedoch auch den Nutzen eines solchen Ansatzes in komplexen Abhängigkeitsnetzwerken. Mittels mathematischer Verfahren kann bestimmt werden, ob ein Austausch auf der Grundlage der bisher ermittelten Interaktionen überhaupt möglich gewesen ist. Das Ergebnis solcher Überlegungen ist damit in der Regel der Ausgangspunkt für weitere Hypothesen und damit für weitere Detailstudien mit dem Ziel der Beantwortung der Frage, ob ein im Netzwerk angelegter möglicher Kontakt tatsächlich realisiert wurde. An zwei einfachen Beispielen soll dies verdeutlicht werden.

10.3.4 Erste Eindrücke des sozialen Netzwerkes: Christoph Clavius als Autor

Bisher sind uns nur wenige Beziehungen zwischen den einzelnen Akteuren explizit bekannt. Trotzdem weist an dieser Stelle die Visualisierung des Netzwerkes ‚basierend auf nur wenigen Beziehungen, auf Auffälligkeiten der Verbreitungsgeschichte hin. Als einfaches Fallbeispiel betrachten wir die Rolle der Ausgaben der Sphaera, bei denen Christoph Clavius[8] als Koautor genannt wird.[9] Eine Anordnung des Netzwerkes gruppiert nach Publikationsorten zeigt, wie sich lokale Schwerpunkte herausbilden. Sacroboscos Sphaera ist hierbei sowohl in den katholischen Regionen als auch in den protestantischen vertreten. Schwerpunkte sind die jeweiligen akademischen Zentren der Konfessionen: auf der protestantischen Seite finden wir Leipzig und Wittenberg und auf der katholischen Leiden, Paris und die oberitalienischen Städte (Abb. 10.3). Sehen wir auf das Subnetz der Ausgaben mit dem Autor Clavius (blaue Quadrate), springt jedoch eine Auffälligkeit ins Auge: während er erwartungsgemäß in den protestantischen Regionen nicht repräsentiert ist, ist er auch im katholischen Paris nicht vertreten. Eine andere Auffälligkeit ergibt sich, wenn wir uns das kleine Teilnetz der Ausgaben mit Clavius als Autor genauer anschauen (Abb 10.4). Die beiden Kölner Ausgaben (Nr. 1957 und Nr. 1820 in der Abbildung) fallen einerseits dadurch auf, dass sie die einzigen Editionen sind, in denen neben Clavius noch weitere Autoren genannt werden. Sie bilden zusätzlich einen eigenen kleinen Cluster. Beide Ausgaben stammen von der Verlegerfamilie Colinus [126, 26], einem Verlagshaus, das bisher nur wenig Beachtung gefunden hat. Das Auffällige dieser Ausgaben ist, dass sie in der Tat keine Editionen der Sphaera darstellen, sondern lediglich einen zusammenfassenden Überblick liefern.[10]

[8] Christoph Clavius, geb. 25.3.1538, gest. 6.2.1612, war einer der herausragenden Gelehrten der 16. Jahrhunderts. Er lehrte Mathematik am Collegium Romanum des Jesuitenordens [5].

[9] Matteo Valleriani geht in [247, S.451 ff] ausführlich auf die Rolle von Christoph Clavius innerhalb des Netzwerkes ein.

[10] Ohne Zweifel wäre das auch ohne die geschilderten Methoden, durch die manuelle Sichtung der Ausgaben deutlich geworden, trotzdem ist dieses ein Beispiel für eine wesentliche Funktion einer systematischer Analyse, nämlich auf Abweichungen der Norm aufmerksam zu machen.

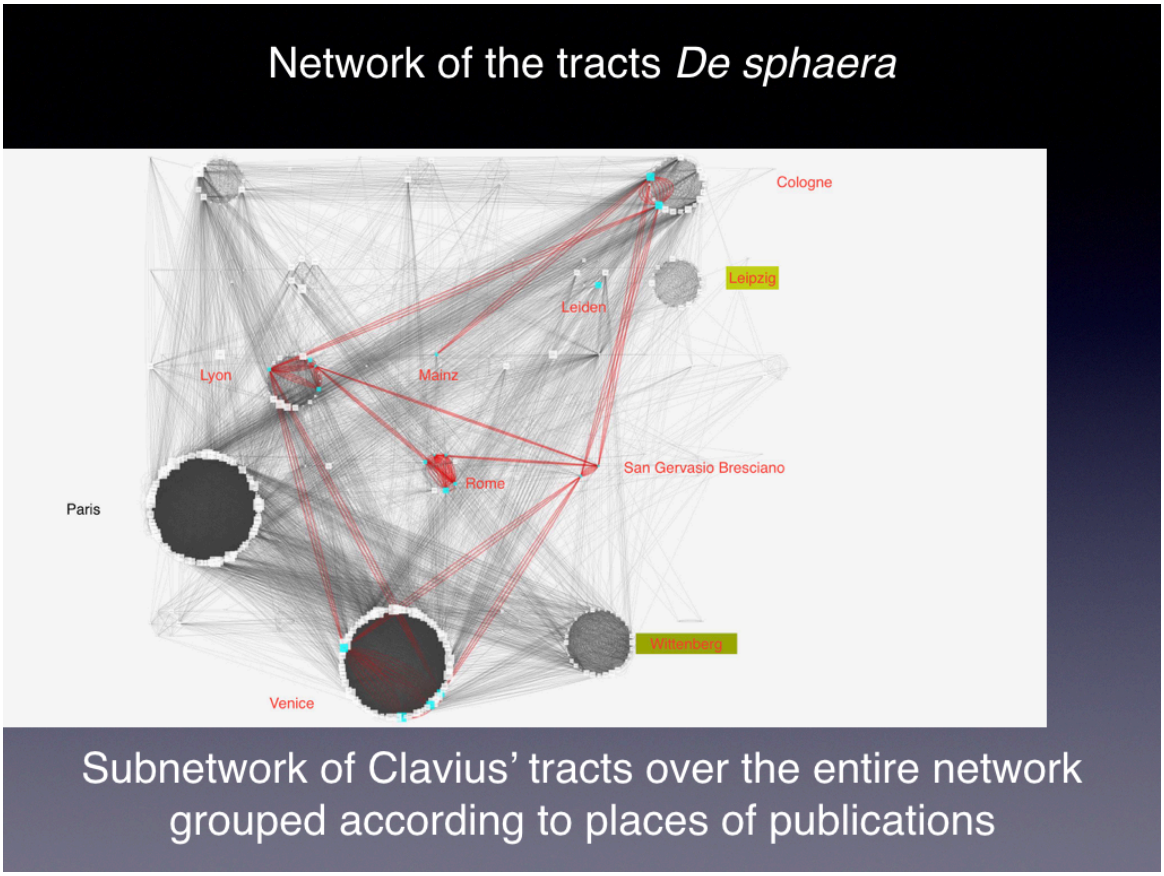


Abbildung 10.3: Gruppierung des Netzwerkes nach Städten, die blauen Punkte markieren Ausgaben mit Clavius als Autor (Stand Dezember 2015, Visualisierung mit Citoscape)

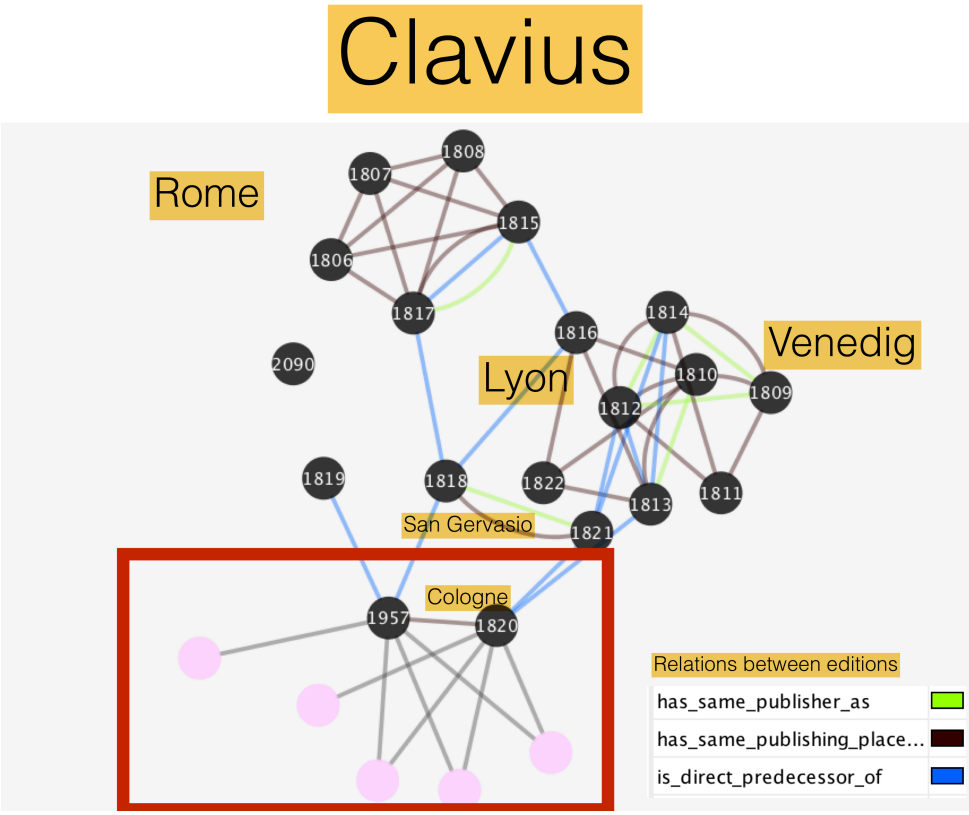


Abbildung 10.4: Teilnetz mit Ausgaben, bei denen Clavius als Autor genannt wird. Die Knoten mit Nummern sind Publikationsereignisse, die anderen Knoten stellen andere Autoren als Clavius dar. (Stand Dezember 2015, Visualisierung mit Visione, Karte von OpenStreetMap)

10.4 Modellierung von internen Strukturen

Wie in der Einführung angesprochen, ist eine unserer zentralen Annahmen, dass die Rekombination der einzelnen Teile der Sphaera, die Aufnahme von neuen Teilen und das Weglassen anderer Teile Hinweise auf die Veränderung des Wissenssystems der Sphaera zulässt. Im Konzept des epistemischen Netzwerkes vermuten wir eine Verbindung zwischen den Textblöcken als Teil des *semiotischen Netzes* und dem *semantischen Netzwerk*. Bei über 300 Editionen der Sphaera stellt sich die Frage, wie eine solche Strukturveränderung analysiert werden kann. Der erste Ansatz ist die Erfassung und genaue Beschreibung des Inhaltes in der Form von auswertbaren Einzeleinträgen. Im Datenmodell auf der Ebene von FRBroo ist dies durch die interne Beziehung der Teile zum gesamten Dokument vorgesehen. Das kann im Wesentlichen durch manuelles Sichten erfolgen. Zur Beschleunigung des Verfahrens wurde hierbei am Institut eine in *metaphactory*[11] eingebettete Annotationsumgebung umgesetzt.

Die größere Herausforderung ist es jedoch, auf die Texte selbst zurückzugreifen. Die ersten Ansätze bestanden in dem Versuch, mittels OCR zumindest einzelne Merkmale zu erschließen. Die Qualität des OCR erzeilt durch *OCRopus*[12] und Tesseract,[13] sowie zum Test auch mit Abbyy Recognition Server 4,[14] erwiesen sich jedoch für diesen Zweck als nicht ausreichend. Zurzeit wird daher nun experimentell der Ansatz verfolgt, auf Layoutmerkmale zurückzugreifen, um Verschiebungen und Veränderungen zu erkennen.

Ziel der Modellierung ist es, die innere logische Struktur der einzelnen Texte zu erfassen und Zusammenhänge zwischen unterschiedlichen Editionen aufzuzeigen. Hierbei sollen sowohl die Argumentationsstruktur als auch der mathematisch-naturwissenschaftliche Inhalt erschlossen werden. Die Argumentationsstruktur umfasst hierbei Elemente, wie sie aus der Diskursanalyse bekannt sind, sowie Strukturelemente, wie sie für die mathematisch-naturwissenschaftliche Literatur der Frühen Neuzeit und Neuzeit typisch werden, also Sätze, Beweise, Hypothesen, Lemmata und Schlüsse. Im Sinne der Netzwerktheorie epistemischer Systeme sind dies die Bausteine des formalen semiotischen Systems der Texte. Dieses korrespondiert mit der semantischen Struktur des damit vermittelten Wissenssystems der Sphaera. Zurzeit liegt ausformuliert hier nur wie oben geschildert eine Verbindung zwischen ganzen Teilen des Textes und Großthemenbereichen vor. In einer nächsten Phase gilt es, jedes Einzelelement zu beschreiben und in die entsprechenden Kontexte zu setzen.

Alternativ zu OCR ist ein vielversprechenderer Ansatz, anstelle des eigentlichen Textes die Einbettung von Diagrammen und Zeichnungen sowie deren Veränderungen zu untersuchen. Bildvergleiche sind mit Open-Source-Paketen wie etwa *TensorFlow* [233] oder mit Hillfe von *OpenCV* [174] mittlerweile auf Standardrechnern möglich.[15]. Die Reihung und der Kontext von Bildern lässt sich dann

[11] Siehe Abschnitt 6.6.

[12] OCRopus ist ein wesentlich von Thomas Breul entwickeltes Paket von Pythonprogrammen zum OCR, das auf dem Einsatz neuronaler Netze beruht [33]. Eine gutes Tutorial von Uwe Springmann und vereinfachende Skripte von David Kaumanns finden sich zusammengefasst unter dem Titel Orcocis in [172].

[13] Tesseract ist eine weitverbreitete Open Source OCR Software, die beim Einsatz auf modernen Druckwerken mittlerweile zuverlässige Ergbnisse liefert [234].

[14] Ein hoch performanter kommerzieller OCR Server der Firma Abbyy [1].

[15] Für eine Einführung siehe z.B. [167]

automatisch vergleichen. Auch Veränderungen von Abbildungen lassen sich gut erkennen, so dass darüber eine Reihung möglich wird.

10.5 Netzwerkanalysen

Die vorhandenen Daten lassen es bereits zu, zumindest prinzipiell Vermutungen über die Dynamik und die Zusammenhänge zwischen den einzelnen Ausgaben zu formulieren und im Rahmen der Theorie dynamischer Netze zu testen. Matteo Valleriani hat ausgehend von der Netzwerkstruktur begonnen, erste Annahmen über die Verbreitung des Sphaera in SIENA[16] zu formulieren. Die ersten Ergebnisse sind jedoch bisher nicht aussagekräftig und deuten darauf hin, dass einerseits mehr Daten über die sozialen Akteure notwendig sind, um Größen zu definieren, die dann entweder als Observable oder Eigenschaften eingesetzt werden können, um Hypothesen zu überprüfen. Andererseits gilt es auch hier kritisch zu hinterfragen, ob im Rahmen der von uns hier dargestellten Theorie epistemischer Netze[17] die Anwendung von Methoden der sozialen Netzwerkforschung überhaupt zielführend sein kann, wenn grundlegende soziologische Annahmen, auf denen SIENA beruht, in der historischen Situation entweder nicht zutreffen oder mangels Daten nicht hinreichend beschrieben werden können. Problematisch ist hierbei bereits die Identifikation der Akteure. In den SIENA-Modellen gehen wir von Akteuren aus, die von bewussten und unbewussten Einflüssen gesteuert Entscheidungen treffen. In unseren Modellen ist jedoch bereits die Definition der Akteuren nicht eindeutig: die Verleger spielen zwar eine zentrale Rolle, stehen jedoch lediglich stellvertretend für eine komplexere Struktur von Einflussbeziehungen. Diese müssen in das Modell als Faktoren mit aufgenommen werden. Diese Einschränkungen schließen jedoch den Einsatz von Verfahren der SNA im historischen Kontext nicht aus. Es gilt sich jedoch auch der Einschränkungen bewusst zu sein. Verhalten sich etwa Knoten des Netzwerkes wie Akteure in den gut untersuchten Netzwerken der Sozialwissenschaften, so kann dies durchaus als Indiz dafür dienen, dass diese Knoten tatsächlich historische Akteure repräsentieren, die sich wie „moderne Akteure“ der sozialen Netzwerkforschung verhalten. Untermauert werden kann dies jedoch nur durch historische Fallstudien.[18]

10.6 Zusammenfassung und Ausblick

Die Arbeiten an der Sphaera zeigen einen Weg auf, wie auch detaillierte Quellenstudien in den theoretischen Rahmen der historischen Netzwerkanalyse eingebettet werden können. Der hier verfolgte Ansatz bettet sowohl das *close reading* von Texten in Form der genauen inhaltlichen Analyse als auch Ansätze eines *distant reading* in die vergleichenden Strukturanalyse basierend auf den Modellen der epistemischen Netzwerktheorie ein. Wir sehen hier jedoch noch einen weiten Weg vor uns. Es müssen die noch vorhandenen technischen Schwierigkeiten im Umgang mit historischen Drucken bewältigt werden, und es bedarf noch eines Modells für eine historisch kritische Strukturanalyse von

[16] Siehe Abschnitt 5.7.

[17] Siehe Abschnitt 5.13

[18] Ausführlicher ist dies in 7.1 diskutiert worden.

Quellen. Wenn auf Volltextanalysen nicht zugegriffen werden kann – sei es wegen der Masse der Texte oder wegen der geringen Qualität des OCR – ist darüber hinaus eine Methodik des *distant reading* noch zu entwickeln, die an die Stelle der Analyse vollständiger Texte treten kann. Die oben erwähnte Bildanalyse kann möglicherweise ein Weg dahin sein.

Kapitel 11

Evaluation der Ergebnisse und Ausblick

In Abschnitt 3.2 hatte ich zentrale Thesen vorgestellt, die den Rahmen für ein Forschungsprogramm zur Entwicklung und Anwendung einer erweiterten Netzwerktheorie für die historische Forschung setzen. Der Anspruch dieser Arbeit war es, diese Thesen aus wissenschaftshistorischer Sicht plausibel zu machen, sowie die zur Verfolgung dieses Ansatzes notwendigen Grundlagen und konkrete Schritte für dessen Realisierung aus Sicht der Informatik vorzustellen. Die einzelnen Kapitel der Arbeit haben sich in unterschiedlicher Weise und Tiefe mit diesen Thesen befasst. Die grundlegende These **HT**, die einen integrierten Ansatz von netzwerktheoretischer und modellierungstheoretischer Beschreibung (NMB) fordert, durchzieht die gesamte Arbeit. Teil I legte die Grundlagen für die Umsetzung dieser These aus Sicht der einer Geschichte der Wissenschaft und des Wissens (Kapitel 2 und 3) sowie aus Sicht der Informatik (Kapitel 4 und 5). Allen Fallbeispielen in Teil II gemeinsam ist, dass sie diese These mit Beispielen aus verschiedenen Anwendungsbereichen untermauern.

Die weiteren Thesen **T1-T3**, die auf den eher beschreibenden Aspekt der NMB abzielen, wurden mit unterschiedlichem Gewicht in den Fallbeispielen behandelt. Es wurde ihre Tragfähigkeit dargestellt. Die Thesen zur Wissensdynamik **WD1-WD4** konnte ich nur anreißen. Indem sie Wissensdynamik mit Netzwerkdynamik verbinden, sind sie jedoch die Voraussetzung, um zu verstehen, warum der geschilderte Weg einer Integration von netzwerktheoretischer und modellierungstheoretischer Beschreibung (NMB), wie sie These **HT** einfordert, aus der Sicht einer historischen Epistemologie (Abschnitt 2.2) überhaupt angegangen wurde.

Ziel der Fallbeispiele war es bewusst, eher auf Breite als auf Tiefe zu setzen, um so das Anwendungsspektrum und die praktische Umsetzung möglichst weit gefächert darzustellen. Dazu wurden Beispiele aus sehr unterschiedlichen Forschungsprojekten eingeführt. Allen Fallbeispielen ist gemeinsam, dass sie die Bedeutung der in 2.3.1 geschilderten Zyklen für Forschungsprozesse und Forschungsdaten für den Umgang mit Daten innerhalb historischer Projekte verdeutlichen. Die Fallstudie zur *Allgemeinen Relativitätstheorie* (Kapitel 7) hatte hierbei ihren Schwerpunkt auf dem Übergang von einer klassischen Datenbank zu einem Modell, das dann der Netzwerkanalyse zugänglich ist. Beleuchtet haben wir hier besonders die Probleme, die sich aus fehlenden Daten ergeben sowie den Schwierigkeiten bei der Interpretation von Netzwerken, die sich unter sich stark verändernden Rahmenbedingungen entwickeln. Es wurde gezeigt, wie Modellierung und Analyse hier dabei helfen, Ordnungsstrukturen übersichtlich darzustellen und neue Strukturen zu erkennen. Die Teilstudie

Strukturen und Netzwerke (Kapitel 8) erweiterte diesen Aspekt um das Problem der Modellierung und Analyse inner-institutioneller Beziehungen und zeigte, wie sich aus der veränderten Sicht auf die Daten neue Fragestellungen ergeben und damit neue Methoden eingeführt werden mussten. Insbesondere galt dieses für die Interpretation der thematischen Cluster im Vergleich zu Ergebnissen der Netzwerktheorie. Die Fallstudie zum *Bau der Domkuppel in Florenz* (Kapitel 9) hatte einen anderen Schwerpunkt, hier wurde die Stärke des Modellierungsansatzes für die Interoperabilität und Anschlussfähigkeit historischer Datenbestände deutlich. Insbesondere wurde die sich durch die Modellierung auftuende Möglichkeit einer mehrschichtigen Datenorganisation nach unterschiedlichen inhaltlichen und formalen Kriterien gezeigt und damit eine wesentliche Voraussetzung für die Verschränkung von semiotischem und semantischem Netzwerk erfüllt, wie sie in der Einführung (Kapitel 1) und ausführlicher in Abschnitt 5.13 dargestellt wurde. Neue Ansätze eröffnen sich durch die Modellierung unter anderem auf Grund der dadurch erleichterten Anbindung externer Datenbestände – wie etwa semantischer (z.B. mittels *WordNet*) oder vertiefender Informationen (z.B. über *Wikidata*). Die frühe Arbeit zur *Sphaera* in Kapitel 10 dagegen setzte den Schwerpunkt auf der Verbindung von semiotischem und sozialem Netzwerk, die ebenfalls in Kapitel 1 diskutiert wurde. An diesem Beispiel wird besonders die Bedeutung eines forschungsgeleiteten Vorgehens für die Datenmodellierung deutlich.

11.1 Ergebnisse aus Sicht der Umsetzung in Softwareumgebungen

Als Ergebnis der in der Arbeit geschilderten Fallstudien liegt ein ausbaubares zu CRM kompatibles Datenmodell für die Beschreibung von Kooperationsformen in der wissenschaftlichen Forschung vor. Konkret hilft dieses, die bisher unverbundenen Daten der beiden Fallstudien zur ART und zur Geschichte der MPG zu verbinden und Fragen an die Daten zu stellen, die über die Entstehungskontexte der Daten hinausgehen. Für die Projekte zur Florentiner Kuppel und der Sphaera konnte beispielhaft gezeigt werden, dass ein Übergang von einem forschungsgeleitet entwickelten Datenbanksystem zu einem CRM-kompatiblen Modell mit begrenztem Aufwand möglich ist. Es konnte damit gezeigt werden, dass sich CRM über den ursprünglichen Kontext zur Verwaltung von Museumsdaten auch für die semantische Modellierung von Forschungsdaten nutzen läßt.

Die im *Triplestore* abgelegten und mittels *metaphactory* darstellbaren Daten lassen sich flexibel aufbereiten. In den Fallbeispielen zeigte sich, dass die Anpassung der Abfragen und Darstellungsformen der Ergebnisse in kurzen Zyklen, die durch sich verändernde Forschungsfragen bestimmt sind, ohne größeren Aufwand möglich ist.

Die Verbindung der Datenbestände mit Auswertungsumgebungen in Form von Pythonskripten hat sich in allen Fallbeispielen als ein flexibles Hilfsmittel erwiesen, auch komplexere Analysen so umzusetzen, dass sie auch an eher qualitativ arbeitende Forscherinnen und Forscher vermittelt werden können. Auch wenn das Erlernen einer Progammiersprache immer noch zumindest in Grundzügen notwendig ist, um eigene Anwendungen zu erstellen, sind die mittels Notizbüchern erzielten und dokumentierten Ergebnisse jedoch auch ohne diese vermittelbar.

Wir sehen insbesondere, dass die Benutzung von Pythonnotizbüchern dabei helfen kann, zeitauf-

wändige Entwicklungen von spezifischen und häufig nur von wenigen genutzten Sonderfunktionen in Oberflächen für Datenbanken zu minimieren.

Mit einfachen Hilfsmitteln ist darüber hinaus der Export von Daten aus dem strukturierten Datenpool des Triplestore in Formate, die dann netzwerkanalytischen Untersuchungen offen stehen, möglich. Diese Exporte können flexibel an Forschungsfragen angepasst werden. Die Daten können dann mit sehr unterschiedlichen Anwendungen analysiert werden. Dadurch kann für spezifische Auswertungen leicht auf existierende Programme, wie etwa Gephi, Visone, Cytoscape oder auch einzelnen R-Pakete, zurückgegriffen werden, die ursprünglich für vollständig andere Anwendungskontexte entwickelt wurden.

Die semantische Modellierung der Daten ermöglicht es hierbei, die inhaltliche Interpretation der Daten weiterhin im Fokus zu halten. Die Gefahr einer impliziten Uminterpretation von Daten dadurch, dass sie mit Hilfsmitteln analysiert werden, in denen (häufig unbewusst für den Nutzer) Annahmen kodiert sind, die für den eigentlichen Anwendungsfall nicht zutreffen, wird hiermit zumindest verringert. Ein Beispiel dafür ist die dargestellte Anwendung der Methoden aus der sozialen Netzwerkanalyse (SNA) auf historische Fragestellungen.

Zusammenfassend zeigt sich, dass mit den jetzt bereits bestehenden technischen Möglichkeiten auch anspruchsvollere Analysen mit einem begrenzten Zeitaufwand für die Programmierung möglich sind, so dass Softwareentwicklung, Entwicklung vom Algorithmen und historisch-qualitative Forschungszyklen besser synchronisiert werden können.

11.2 Folgerungen für die Anwendung quantitativer Methoden als heuristische Hilfsmittel

Die systematische Zerlegung der einzelnen Aspekte der Datenerzeugung, -verwaltung und -analyse nach pragmatischen Gesichtspunkten ist hierbei, wie die Fallstudien zeigen, notwendige Voraussetzung. Sie ermöglicht die Diskussion über Strukturen, die normalerweise in großen Datenbankprojekten in den Geschichtswissenschaften nur im Hintergrund stehen. Eine Reflexion über Datenstrukturen wird so in einem interdisziplinären Kontext möglich, da die Auswirkungen von fehlenden Daten und unscharfen Beschreibungen sowie von unterschiedlichen Auswertungsmethoden hier sowohl für die historisch arbeitenden Wissenschaftlerinnen und Wissenschaftler als auch für die aus dem Bereich der Informatik kommenden Forscherinnen und Forscher transparent werden.

Die vorhergehenden Kapitel haben deutlich gemacht, dass es auf Grundlage der unterschiedlichen Themenfelder, der heterogenen Quellenlage und schließlich der hohen Interdisziplinarität der Forschung in den Geisteswissenschaften, für die die Wissenschaftsgeschichte exemplarisch steht, keine monolithische Lösung geben kann, die für alle Anwendungsszenarien optimal von Seiten der Softwareanbieter angepasst werden kann. Die dynamische Veränderung der Fragestellung mit dem Fortschritt in der Forschung bedeutet auch für die Anwendungssoftware in den Geisteswissenschaften ständige Anpassungen. Die für zumindest weite Bereiche der Naturwissenschaften übliche enge Verzahnung von Programmentwicklung und Forschung ist auch für die Geisteswissenschaften in Zukunft unumgänglich. Dies hat grundlegende Konsequenzen, insbesondere für die Gestaltung von

Ausbildungsgängen, von der Schule bis hin zur universitären Ausbildung. Auch für die Geisteswissenschaften müssen von Beginn an die neuen Kulturtechniken, die mit der Informationsrevolution verbunden sind, Teil der Methodenvermittlung sein. Die in dieser Arbeit geschilderten Ansätze, insbesondere die Verbindung der Entwicklung einzelner Skripte durch die Wissenschaftler selbst, wie es beispielhaft durch den Einsatz von Pythonnotizbüchern möglich ist, ist hier ein vielversprechender Ansatz, der auch in die unterschiedlichen Curricula eingehen könnte.

Die ersten Beispiele aus dem Bereich der Allgemeinen Relativitätstheorie sowie dem Projekt zur Geschichte der Max-Planck-Gesellschaft zeigen die Möglichkeiten des quantitativen Herangehens auch an historische Probleme. Es gilt, dass diese Methoden zunächst sicherlich nicht mehr sind als zusätzliche heuristische Hilfsmittel, etwa im Sinne eines klassischen Findmittels für Quellen – aber auch nicht weniger. Quantitatives Arbeiten und detaillierte Fallstudien sind hierbei keine Gegensätze, sondern zwei Seiten einer Medaille. Ohne Fallstudien ist die Interpretation quantitativer Daten nicht möglich, da ihre Rückbindung an die Realität stets über die Fallstudie führt –- umgekehrt können quantitative Daten vor unzulässiger Verallgemeinerung schützen.

Wir haben auch gesehen, dass auf der methodischen Seite noch eine ganze Reihe von Desideraten bestehen, die es aufzufüllen gilt. Methoden aus den Sozialwissenschaften lassen sich nicht direkt in die historische Forschung übertragen; so fehlen überzeugende Modelle, die auch über Umbruchphasen Aussagen zulassen und Veränderungen von Netzwerken auch im historischen Kontext verlässlich vorhersagen. Die vorhandenen Hilfsmittel zur Analyse von Netzwerken und insbesondere ihre Visualisierung sind jedoch auch heute bereits ausreichend, um diese auch als heuristische Werkzeuge in der alltäglichen Arbeit des Historikers oder der Historikerin zu nutzen. Die Transformation der Ausgangsdaten in Formate, die der Netzwerkanalyse zugänglich sind, ist jedoch immer noch aufwendig. Der vorgestellte Ansatz, aus SPARQL direkt Netzwerke zu erzeugen, ist aus meiner Sicht ein erster wesentlicher Schritt in diese Richtung. Hier müssen jedoch noch viele Schritte folgen.

In der Fallstudie zur Geschichte der MPG bekommen wir zwar durch das Experimentieren mit unterschiedlichen Netzwerkkonfigurationen Hinweise auf mögliche weitere Strukturen, die es zu untersuchen gilt. Der jetzige Stand zeigt aber auch die Problemlage auf: Die Quellen, die uns vorliegen, zeigen Interaktionen durch eine sehr enge Brille, nämlich das Handeln von Personen in Gruppen, die durch die Institution vorgegeben sind. Wir sehen in begrenztem Maße auch die Veränderung der Institutionen durch die handelnden Personen. Es fehlt uns noch ein Angriffspunkt, um die sozialen Netzwerke mit semantischen und semiotischen Netzwerken zu verbinden. Ich komme unten noch einmal auf das prinzipielle Verhältnis dieser Strukturen zurück. Einen Ansatz sehe ich in der Analyse der Investitionen in Großgeräte und Rechenanlagen sowie der Analyse der Publikationen der Hauptakteure im Netzwerk. Hier liegt es methodisch nahe, sich von der näheren Vergangenheit schrittweise weiter zurück zu bewegen, da zumindest ab etwa den 1990er Jahren der Zugang zu Publikationsdaten, Volltexten und Abstracts deutlich einfacher ist und diese damit rechnergestützt ausgewertet werden können. Für diese Zeit sollten auch Daten auffindbar sein, wie etwa Zeitungsarchive und Finanzdaten, die es erlauben, zumindest Hinweise auf die Einbettung der internen Entwicklung in einen breiteren gesellschaftlichen Kontext zu gewinnen.

Alle Fallbeispiele machen jedoch das große Defizit deutlich, das trotz aller Digitalisierungsbemü-

hungen immer noch besteht. Daten sind in den seltensten Fällen in einer tatsächlich interoperablen Form bereitgestellt. Methoden zum Identifizieren von Entitäten sind immer noch mit großen Fehlern behaftet, so dass selbst bei vorhandenen Daten, wie den Publikationsdaten auf der einen Seite und den Mitgliederverzeichnissen im Projekt zur Geschichte der MPG auf der anderen Seite das Zusammenbringen dieser Daten immer noch mit großem Aufwand und hohen Fehlerraten verbunden ist. Fragestellungen, die unmittelbar auf der Hand liegen und sich theoretisch mit den vorhandenen Quellen auch mit digitaler Unterstützung lösen lassen, lassen sich oft doch nicht angehen, weil die entsprechenden Methoden noch nicht ausgereift genug sind und auch – das ist ein zentraler Punkt – die dahinter stehenden Prinzipien nicht immer für alle Beteiligten klar sind. Wir haben diesen Punkt in der Arbeit nur anreißen können. Am deutlichsten ist dies im Bereich der Anwendung netzwerktheoretischer Methoden. Ergebnisse lassen sich zwar mit den aus den Sozialwissenschaften bekannten und etablierten Werkzeugen erzielen. Inwieweit diese jedoch methodisch wirklich tragfähig sind, ist weit schwieriger zu beurteilen. In den ausgereiften Algorithmen, etwa für ERGM oder SIENA, sind sozialwissenschaftlich fundierte Handlungsannahmen tief verankert und es ist nötig, diese noch einmal sehr genau im historischen Kontext zu überprüfen. Wie jedoch in der Arbeit häufiger betont, ist diese mögliche Ungenauigkeit kein Grund, diese Methoden nicht zu bemühen. Sie haben eine zweifache Berechtigung. So haben wir gezeigt, dass die Arbeit mit großen Datenmengen dadurch tatsächlich vereinfacht wird und wir Hypothesen erst auf Grund dieser Ergebnisse erzeugen können. Andererseits stehen die Geisteswissenschaften vor der fundamentalen Herausforderung, diese Techniken beherrschen zu müssen. Wie in den einleitenden Kapiteln dieser Arbeit besprochen, ist dies unumgänglich, wenn die Geisteswissenschaften im digitalen, vernetzten Zeitalter bestehen wollen. Zur Funktion der Geisteswissenschaften, gesellschaftliche Prozesse kritisch zu reflektieren, gibt es keine Alternative. Geisteswissenschaften können das jedoch nicht leisten, wenn sie sich nicht selbst auch die entsprechenden Techniken aneignen, um die Grundprinzipien der Wissensorganisation im Netzwerkzeitalter zu verstehen.

Die Arbeit zeigt, dass dies kein Hexenwerk ist, sondern bereits durch den Einsatz von bestehenden Methoden Fortschritte erzielt werden können. Ohne Kreativität und eine gewisse Risikobereitschaft werden sich die Geisteswissenschaften diese Methodik nicht aneignen. Die Aufnahme von Methoden aus der Informatik in den geisteswissenschaftlichen Fächerkanon, wie z. B. die Simulation, gibt der Geisteswissenschaft eine Experimentalkomponente. Das sollte in dieser Arbeit deutlich geworden sein. In den Geisteswissenschaften sind Experimente jedoch immer noch ein eher ungewöhnlicher Weg. Die Geschichtswissenschaften, und hier insbesondere auch die Wissenschaftsgeschichte, sind in diesem Kontext gute Mediatoren, da die Rekonstruktion historischer Experimente und der Bedingungen von Forschung zumindest im Kontext der historischen Epistemologie ein durchaus akzeptierter Teil des Baukastens der Forscherinnen und Forscher ist.

Neben dem Experimentieren ist der modellierungstheoretische Ansatz die andere Seite der Medaille. Semantische Modellierung ist ein Schritt in die Richtung eines einfacheren Zusammenführens von Datenbeständen. Sie erlaubt nicht nur logische Operationen wie die Überprüfung auf Vollständigkeit und Widersprüche oder die Klassifikation per Inferenz, sondern – und hier liegt die entscheidende Stärke – sie ermöglicht es, eine Sprache zu entwickeln, die der vernetzen Wissensstruktur angepasst

ist.

Insbesondere besitzen wir mit den vorgestellten Grundlagen nun ein Handwerkszeug, um eine komplexere Koevolutionstheorie wissenschaftlicher Entwicklungen anzugehen. Diese Arbeit hat hierbei bewusst zunächst auf die Grundlagen beider theoretischer Ansätze der mathematischen Netzwerktheorie und der Modellierungstheorie gesetzt, um zunächst einmal auszuloten, wie tragfähig diese sind und dann darauf aufbauend eine erweiterte Theorie entwickeln zu können. Im nächsten Schritt geht es nun darum, die entwickelten Modelle systematisch zu erweitern. Grundsätzlich geht es darum, Parallelentwicklungen der in der Einleitung geschilderten unterschiedlichen Netzwerke des epistemischen Systems nun exakter zu beschreiben und genauere Hypothesen für die Bedingungen ihrer Koevolution zu entwickeln. Dazu wird es komplexerer mathematischer Methoden bedürfen als derjenigen, die im Rahmen dieser Arbeit eingeführt wurden. Der Ansatz, den wir in der Zukunft aufbauend auf den Erfahrungen der geschilderten Projekte gehen werden, lässt sich zusammenfassend so umreißen.

Auf der einen Seite bedarf es einer Beschreibung und Kodierung der Entwicklung spezifischer Forschungsfelder. Im Beispiel der ART sind wir hier dem Ansatz gefolgt, das soziale System der Kontakte der Forscher untereinander und deren Entwicklungsdynamik mit Hilfe der Netzwerktheorie zu beschreiben. Mittels bibliometrischer Methoden verfolgten wir die Entwicklung des in unserer Terminologie semiotischen Systems. Die darauf aufbauenden Untersuchungen – in diesem Falle mittels Citespace – geben erste Hinweise auf die dahinter liegende Entwicklung von Kategorien und Begriffen und damit auf das semantische System. Bei der Untersuchung der Kommissionen liegt der Schwerpunkt zunächst auf dem sozialen System, jedoch hier mit der besonderen Beachtung der Ausprägung von Strukturen, die mittels der Arbeitshypothese der Cluster aufgrund inhaltlicher Nähe entstehen sollten. Auch dies ist ein Beispiel für die Kopplung der Bildung von Forschungsschwerpunkten, sozialer Struktur und deren Institutionalisierung. Im Beispiel des Duomo-Projekts haben wir prinzipiell gezeigt, dass eine Überführung komplexer Datenstrukturen in ein Format, das dann der netzwerktheoretischen Auswertung offen steht, in stringenter Art und Weise möglich ist. Wir zeigen hier auch den Zusammenhang zwischen der Modellierung der Interpretation von historischen Daten auf der einen Seite und der Abbildung der Quellenstruktur auf der anderen Seite, ohne die historische Anwendungen nicht denkbar sind.

11.3 Ausblick

Während wir die Konsequenzen der Thesen **HT** und **T1-3** als Grundlage einer NMB in dieser Arbeit verdeutlicht haben, ist die Umsetzung der Thesen **WD1-WD4** von theoretisch-qualitativen Überlegungen zu einer Wissensdynamik hin zu einem mathematischen Modell noch ein Desiderat. Dass der methodische Ansatz einer Koevolution der unterschiedlichen Teilsysteme des epistemischen Netzwerkes ein vielversprechender Ansatz ist, zeigen die in dieser Arbeit geschilderten Experimente mit historischen Daten. Noch aber liegt ein weiter Weg hin zu Computational Humanities vor uns. Zur Weiterentwicklung dieses Ansatzes ist es notwendig, auf der einen Seite die modellierungstheoretische und damit semantische Struktur präziser zu beschreiben und zugleich auf der anderen Sei-

te die netzwerktheoretische Multilevelanalyse weiter zu verfeinern. Wir sehen, dass besonders die Umbruchphasen, also gerade die Teile der Struktur, die von besonderem Interesse sind, in der Netzwerkanalyse nur schwer mit den herkömmlichen Methoden aus der sozialen Netzwerkanalyse zu behandeln sind. Die Phasen zwischen den Umbrüchen lassen sich dagegen in der Regel gut mit Modellen erfassen. Umgekehrt heißt dies, dass der Kollaps eines Modells gerade ein Indikator für solche Umbruchphasen ist. Abschließend möchte ich den Ausblick daher mit einer Beschreibung des Modells beenden, das wir in Zukunft zu behandeln haben.

Auf der Modellierungsseite steht das folgende Modell: Wir nehmen an, dass zentrale Ereignisse dazu beitragen, dass sich Strukturen in den drei Netzwerken verändern. In unseren Fallstudien können dies neue Beobachtungen, Konferenzen oder auch neue Methoden sein. Diese Ereignisse betreffen immer das gesamte epistemische Netz. Diese Struktur gilt es so genau wie möglich zu beschreiben. Sie gibt den epistemischen Handlungsraum vor. Dieser hat die zentrale Funktion, einen Raum mit verminderter Komplexität und endlicher Dimension zu definieren, der dann mit mathematischen Methoden der Netzwerkanalyse behandelt werden kann. Diese erlauben, zumindest im Rahmen der heutigen Theorien, nur ein sehr eng begrenztes Maß an Unsicherheit als externem Parameter. Andererseits zeigt der Kollaps der Modelle Defizite in der modelltheoretischen Beschreibung auf.

Auf der Netzwerkseite des Modells sehen wir die drei Netzwerke im Sinne der Multileveltheorie von Netzwerken als einzelne Ebenen. Auf diesen Ebenen bestehen Bindungen auf der sozialen Seite über persönliche Kontakte, auf der semiotischen Seite stehen Methoden, Formalismen, aber auch Instrumente miteinander in Wechselwirkung. So kann etwa ein notwendiges Großgerät oder ein Algorithmus unterschiedliche Systeme miteinander verbinden. Auf der semantischen Seite stehen miteinander verbundene Konzepte, etwa im Sinne von Einfluss-, Teil-Ganzes-Beziehungen oder in Form von Widersprüchen. Diese bestimmen die Kanten des Netzwerkes und ergeben sich unmittelbar durch die Projektion der semantischen Modellierung auf das Netzwerk.

Das Wesentliche wird nun sein, Handlungsmodelle zu entwickeln, die die von Ereignissen angestossenen Möglichkeiten der Veränderung der Netzwerke genauer beschreiben und dann der mathematischen Modellierung offenstehen – zum Beispiel im Sinne von Kosten- und Korrelationsfunktionen auf den Netzwerken.

Dies sei für ein einfaches Beispiel geschildert. Wir nehmen als Ausgangspunkt den Zustand des Netzwerkes vor einem möglichen zentralen Ereignis wie einer bedeutenden Konferenz. Einzelne Personen sind bereits durch Links miteinander verbunden. Diese Personen stehen außerdem in Verbindung zu experimentellen Methoden sowie theoretischen Konzepten. Wir nehmen nun an, dass durch Ereignisse auf der Konferenz unterschiedliche neue, zunächst wahrscheinlich schwache Bindungen entstehen. Dies kann dadurch geschehen, dass indirekte Beziehungen, die zwischen Personen entweder über das semiotische oder das semantische Netz bereits bestehen, von den Akteuren realisiert werden und sich damit zu Bindungen im sozialen Netzwerk ausprägen. Umgekehrt kann ein Ereignis wie der Bau eines Großgerätes, etwa eines Beschleunigers, dazu führen, dass über das Gerät methodische Ansätze zusammengeführt werden, die bisher nicht miteinander in Kontakt standen. Im Netzwerk wird dadurch das Potenzial angelegt, das zu weiteren Verbindungen auch auf sozialer und

semantischer Ebene führen kann, die jedoch von den Akteuren noch nicht realisiert werden.

Dieses Koevolutionsmodell erlaubt einerseits eine systematische Einordnung von Handlungen und Verfahren auf der modellierungstheoretischen Seite im Sinne einer wissenschaftstheoretischen Erklärung, andererseits gibt es eine Richtlinie für die Rekonstruktion von historischen Kontexten. Soziale Netzwerke können im historischen Rückblick nur auf der Grundlage vorgefundener Artefakte, das heißt also auf der Grundlage unserer Rekonstruktion des semiotischen Netzwerkes, eingegrenzt werden. In diese gehen Modelle wie das Verhältnis von implizitem und explizitem Wissen oder Theorien über die Transmission von Wissen ein, die auf der semantischen Ebene liegen. Auf diese Art entsteht innerhalb des Systems jeweils ein unmittelbarer Rückbindungseffekt des aktuellen Wissenssystems der Historiker und Wissenschaftstheoretiker an das zurückliegende System, das es zu beschreiben gilt. Netzwerktheoretisch entsteht so ein komplexes Multilevelnetzwerk mit einer komplexen Zeitentwicklung, die zumindest dem Ziele nach im Sinne einer Evolutionstheorie einen gerichteten Zeitpfeil beinhaltet.

Unsere bisherigen Arbeiten geben zumindest Anlass dafür, zu glauben, dass eine approximative Annäherung an das Problem durch die Bestimmung geeigneter Teilsysteme, wie z. B. der Entwicklung eines Forschungsfeldes, möglich ist. Strukturell wird dies dadurch möglich, dass wir externe Faktoren als Parameter an das System übergeben und es bis auf diese Parameter als unabhängig betrachten. Die mathematische Herausforderung besteht dann darin, die Fehler abzuschätzen, die dadurch in quantitativen Auswertungen entstehen. Die modellierungstheoretische Herausforderung ist, das System weiterzuentwickeln, so dass die Fehler minimiert werden. Dies ist mit der Hoffnung verbunden, dass mathematische Modelle zeigen können, dass bestimmte Einzelentwicklungen nicht bekannt sein müssen, um hinreichend gute Erklärungen zu liefern – analog dazu, dass Aussagen über Gesamtsysteme, wie etwa das Wahlverhalten oder Systeme in der statistischen Physik, häufig präzise möglich sind, ohne den Einzelfall zu analysieren. Umgekehrt sehen wir die Probleme dieser Methodik etwa in den letzten Jahren bei den häufiger nicht zutreffenden Vorhersagen von Wahlergebnissen insbesondere für kleinere Gruppen, wenn Umbrüche im System strukturelle Veränderungen der Modelle verlangen.

Neben den immer noch bestehenden grundlegenden methodischen Problemen ist vor allem auch deutlich, dass die Etablierung angewandter *Computational Humanities* eine notwendige Voraussetzung für die Weiterentwicklung des Feldes ist. Analog zu anderen anwendungsorientierten Forschungen sollten hier sowohl Methoden als auch technische Kapazitäten bereitgestellt werden, damit die Methoden auch in der praktischen Forschung und in nennenswertem Umfang angewandt werden können.

Teil III

Anhang

Kapitel 12

Beispiele für SPARQL-Queries

12.1 Strukturen und Netzwerke

Im folgenden einzelne Beispiele für SPARQL-Abfragen zum Kapitel „8 Strukturen und Netzwerke“.

```
prefix sr: <http://ontologies.mpiwg-berlin.mpg.de/scholarlyRelations/>
prefix rdfg: <http://www.w3.org/2004/03/trix/rdfg-1/>
prefix gmpgg: <http://ontologies.mpiwg-berlin.mpg.de/gmpg/graph/>
insert data {graph gmpgg:personsCommissionStage1
              {
              gmpgg:persons rdfg:subGraphOf  gmpgg:personsCommissionStage1;
                                                          rdf:type rdfg:Graph.
              gmpgg:commissions rdfg:subGraphOf  gmpgg:personsCommissionStage1;
                                                          rdf:type rdfg:Graph.


  gmpgg:commissionType rdfg:subGraphOf gmpgg:personsCommissionStage1;
                                                             rdf:type rdfg:Graph.
              gmpgg:resourceLinks rdfg:subGraphOf gmpgg:personsCommissionStage1;
                                                           rdf:type rdfg:Graph.
  gmpgg:calculatedYears rdfg:subGraphOf gmpgg:personsCommissionStage1;
                                                          rdf:type rdfg:Graph.
     gmpgg:personsCommissionStage1 rdf:type rdfg:Graph.
     gmpgg:personsCommissionStage1 rdf:type sr:GraphSet.
     }
   }
```

Abbildung 12.1: Graphensatz mit allen Informationen aus der Datenbank (siehe 8.3.1)

```
prefix sr: <http://ontologies.mpiwg-berlin.mpg.de/scholarlyRelations/>
prefix rdfg: <http://www.w3.org/2004/03/trix/rdfg-1/>
prefix gmpgg: <http://ontologies.mpiwg-berlin.mpg.de/gmpg/graph/>
insert data {graph gmpgg:personsCommissionStage2
             {
             gmpgg:persons rdfg:subGraphOf  gmpgg:personsCommissionStage2;
                                                          rdf:type rdfg:Graph.
             gmpgg:commissions rdfg:subGraphOf  gmpgg:personsCommissionStage2;
                                                          rdf:type rdfg:Graph.

  gmpgg:commissionType rdfg:subGraphOf gmpgg:personsCommissionStage2;
                                                             rdf:type rdfg:Graph.
             gmpgg:resourceLinks rdfg:subGraphOf gmpgg:personsCommissionStage2;
                                                           rdf:type rdfg:Graph.
  gmpgg:calculatedYears rdfg:subGraphOf gmpgg:personsCommissionStage2;
                                                          rdf:type rdfg:Graph.

             gmpgg:estimatedTimeSpans rdfg:subGraphOf
              gmpgg:personsCommissionStage2;
              rdf:type rdfg:Graph.
    gmpgg:personsCommissionStage2 rdf:type rdfg:Graph.
    gmpgg:personsCommissionStage2 rdf:type sr:GraphSet.


    }
  }
```

Abbildung 12.2: Graphensatz mit allen Informationen aus der Datenbank II (siehe 8.3.1)

```
prefix sr: <http://ontologies.mpiwg-berlin.mpg.de/scholarlyRelations/>
prefix rdfg: <http://www.w3.org/2004/03/trix/rdfg-1/>
prefix gmpgg: <http://ontologies.mpiwg-berlin.mpg.de/gmpg/graph/>
insert data {graph gmpgg:persons_all
             {
             gmpgg:persons rdfg:subGraphOf  gmpgg:persons_all;
                                                                    rdf:type rdfg:Graph.

    gmpgg:workingdataCommissions rdfg:subGraphOf  gmpgg:persons_all;
                                                                    rdf:type rdfg:Graph.

gmpgg:ressourcelinks rdfg:subGraphOf  gmpgg:persons_all;
       rdf:type rdfg:Graph.

gmpgg:classify_commission_membership rdfg:subGraphOf  gmpgg:persons_all;
       rdf:type rdfg:Graph.
 gmpgg:commissions rdfg:subGraphOf  gmpgg:persons_all;
                                                                    rdf:type rdfg:Graph.

 gmpgg:classify_cluster rdfg:subGraphOf  gmpgg:persons_all;
                                                                    rdf:type rdfg:Graph.

                          gmpgg:functions rdfg:subGraphOf  gmpgg:persons_all;
                                                                    rdf:type rdfg:Graph.
  gmpgg:commissionType rdfg:subGraphOf gmpgg:persons_all;
                                                                        rdf:type rdfg:Graph.
             gmpgg:resourceLinks rdfg:subGraphOf gmpgg:persons_all;
                                                                      rdf:type rdfg:Graph.
  gmpgg:calculatedYears rdfg:subGraphOf gmpgg:persons_all;
                                                                     rdf:type rdfg:Graph.

             gmpgg:personsAdd2 rdfg:subGraphOf gmpgg:persons_all;
                                                                     rdf:type rdfg:Graph.

             gmpgg:personsAdd1 rdfg:subGraphOf gmpgg:persons_all;
                                                                     rdf:type rdfg:Graph.

             gmpgg:classify_commission_membership rdfg:subGraphOf
                  gmpgg:persons_all; rdf:type rdfg:Graph.

             gmpgg:add_cluster rdfg:subGraphOf gmpgg:persons_all;
                                                                     rdf:type rdfg:Graph.

             gmpgg:classify_cluster rdfg:subGraphOf
                            gmpgg:persons_all; rdf:type rdfg:Graph.

             gmpgg:EstimatedTimeSpans rdfg:subGraphOf
              gmpgg:persons_all;
              rdf:type rdfg:Graph.
    gmpgg:persons_all rdf:type rdfg:Graph.
    gmpgg:persons_all rdf:type sr:GraphSet.
    }
  }
```

Abbildung 12.3: Graphensatz mit allen Informationen aus der Datenbank und Wikipedia (siehe 8.3.3)

```
prefix skos: <http://www.w3.org/2004/02/skos/core#>
prefix sr: <http://ontologies.mpiwg-berlin.mpg.de/scholarlyRelations/>
prefix crm: <http://erlangen-crm.org/160714/>
select distinct
?nm__b2 ?nm__e2 ?ns ?ns__label ?nm__begin ?nm__end ?nm__label ?ns_nm__begin ?nm
?ns_nm__end ?nm__type1 ?nm__sec ?ns_nm__type_memb ?ns_nm__type2_memb ?nm__cls
?nm__isAppointmentAWM
where
{
  ?p crm:P12i_was_present_at ?km;
     rdf:type crm:E21_Person;
    sr:has_id ?p_id;
    sr:has_name ?p_name.
  ?p_id rdfs:label ?ns.
  ?p_name rdfs:label ?ns__label.

  ?km rdf:type sr:CommissionMembership;
      sr:is_part_of_commission ?k.

   ?k sr:has_id ?k_id;
      sr:has_name ?k_name.
   ?k sr:belongs_to ?sec.
   ?sec rdfs:label ?nm__sec.
   ?k_id rdfs:label ?nm.
   ?k_name rdfs:label ?nm__label.
optional {
   ?km crm:P4_has_time-span ?ts.
   ?ts sr:begins ?b.
   ?b rdf:type sr:Year;
      rdfs:label ?ns_nm__begin.
   }
   optional {
    ?km crm:P4_has_time-span ?ts.
    ?ts sr:ends ?e.
    ?e rdf:type sr:Year;
       rdfs:label ?ns_nm__end.
   }
   optional {
   ?k crm:P2_has_type ?kt.
   ?kt rdfs:label ?nm__type1.
   }
    optional {
   ?k crm:P2_has_type ?kt_AWM.
   ?kt_AWM rdfs:label "Berufung AWM".
   }
   ?km crm:P2_has_type ?kmt.
   ?kmt rdfs:label ?ns_nm__type_memb.
optional{
   ?kmb skos:narrower ?kmt.
   ?kmb rdfs:label ?ns_nm__type2_memb
optional {
   ?cl crm:P41_classified ?k.
   ?cl crm:P42_assigned ?a.
   ?a rdfs:label ?nm__cls.
   }
optional {
   ?k  crm:P92i_was_brought_into_existence_by ?k_s.
   ?k_s crm:P4_has_time-span ?k_ts_s.
   ?k_ts_s ?xx ?k_ts_sy.
   ?k_ts_sy rdf:type sr:Year;
      rdfs:label ?nm__begin.
  }
optional {
   ?k  crm:P93i_was_taken_out_of_existence_by ?k_e.
   ?k_e crm:P4_has_time-span ?k_ts_e.
   ?k_ts_e ?xx_y ?k_ts_ey.
   ?k_ts_ey rdf:type sr:Year;
      rdfs:label ?nm__end.
  }
optional {
  ?k crm:P4_has_time-span ?tsE.
  ?tsE a sr:Estimated_Time-Span;
             sr:begins/rdfs:label ?e_b;
            sr:ends/rdfs:label ?e_e.
}
bind( if(bound(?nm__begin),?nm__begin,"--") as ?nm__begin2)
bind( if(bound(?nm__end),?nm__end,"--") as ?nm__end2)
bind( if(bound(?nm__begin),?nm__begin,?e_b) as ?nm__b2)
bind( if(bound(?nm__end),?nm__end,?e_e) as ?e_e2a)
bind( if(bound(?e_e2a),?e_e2a,?nm__begin) as ?nm__e2)
bind( if(bound(?kt_AWM),"yes","no") as ?nm__isAppointmentAWM)
}
```

Abbildung 12.4: SPARQL - Query zur Erzeugung des ersten bipartiten Graphen. Die Bezeichnung der Variablen sind so angepasst das *generateBiPartiteGraph* sie verarbeiten kann.

12.2 Die Jahre der Kuppel

Einzelne prototypische Beispiele für Abfragen an die Datenbank zum Bau der Kuppel des Florentiner Doms (Kapitel 9).

```
 prefix duomo: <http://ontologies.mpiwg-berlin.mpg.de/duomo/>
 prefix efrbroo: <http://erlangen-crm.org/efrbroo/>
 prefix ecrm: <http://erlangen-crm.org/current/>
 SELECT ?year (COUNT(distinct ?int) as ?int_n) WHERE {
  ?int rdf:type duomo:Duomo_Event;
     ecrm:P4_has_time-span ?ts.
  optional {
  ?ts duomo:has_XSD_date ?date1
  }
  optional {
  ?ts duomo:has_startDate/duomo:has_XSD_date ?date2
  }
  bind( if(bound(?date1),?date1,?date2) as ?date)
   bind( year(?date) as ?year)
filter (?year < 1900)
 }q
 GROUP BY ?year
 ORDER BY ?year
```

Abbildung 12.5: Zeitliche Ordnung der Ereignisse (siehe 9.3.2)

```
    PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX : <http://ontologies.mpiwg-berlin.mpg.de/duomo/>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix duomo: <http://ontologies.mpiwg-berlin.mpg.de/duomo/>
prefix wn20: <http://www.w3.org/2006/03/wn/wn20/instances/>
prefix wn31: <http://wordnet-rdf.princeton.edu/wn31/>
prefix wnont31: <http://wordnet-rdf.princeton.edu/ontology#>
prefix lemon: <http://lemon-model.net/lemon#>

select distinct ?le ?y where {
            ?cf lemon:writtenRep "metal"@eng.
            ?s lemon:canonicalForm ?cf.

            ?synset wnont31:synset_member ?s;
                    wnont31:part_of_speech wnont31:noun;
                    wnont31:hyponym+ ?refs.
            ?sense  lemon:reference ?refs.
            {
   ?le lemon:sense ?sense.
}
   union
            {
             ?synset wnont31:synset_member ?le.
            }
          ?de duomo:is_related_to_lexicalEntry ?le.
%?de duomo:is_primary_related_to_lexicalEntry ?le.

         ?a ?p ?de.
         select ?p where {?p rdfs:subPropertyOf duomo:has_term_atom.}
   ?pt duomo:has_part_term ?a.

    }
```

Abbildung 12.6: sparql:wordnetTermeDuomot

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX : <http://ontologies.mpiwg-berlin.mpg.de/duomo/>
prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>
prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
prefix duomo: <http://ontologies.mpiwg-berlin.mpg.de/duomo/>
prefix wn20: <http://www.w3.org/2006/03/wn/wn20/instances/>
prefix wn31: <http://wordnet-rdf.princeton.edu/wn31/>
prefix wnont31: <http://wordnet-rdf.princeton.edu/ontology#>
prefix lemon: <http://lemon-model.net/lemon#>
prefix ecrm: <http://erlangen-crm.org/current/>

SELECT ?year (COUNT(distinct ?record) as ?record_n)
where {
{
select distinct ?record ?year  where {
            ?cf lemon:writtenRep "metal"@eng.
            ?s lemon:canonicalForm ?cf.

            ?synset wnont31:synset_member ?s;
                    wnont31:part_of_speech wnont31:noun;
                    wnont31:hyponym+ ?refs.
            ?sense  lemon:reference ?refs.
            {
   ?le lemon:sense ?sense.
}
   union
            {
             ?synset wnont31:synset_member ?le.
            }
          ?de duomo:is_related_to_lexicalEntry ?le.
%?de duomo:is_primary_related_to_lexicalEntry ?le.

            {
              ?a ?p ?de.
              {
              select ?p where {?p rdfs:subPropertyOf duomo:has_term_atom.}
              }
              ?pt duomo:has_part_term* ?a.
              ?pt  a duomo:Term_Root.
              ?reges duomo:has_term_structure ?pt.
              ?ev ecrm:P141_assigned ?reges;
                  ecrm:P140_assigned_attribute_to ?record.
       } union {
         ?st duomo:has_analytical_type ?de.
         ?event duomo:expresses_statement ?st.
         ?record duomo:has_interpretation ?event.
       }
       ?record duomo:has_date_of_writing ?date.
   ?date duomo:has_XSD_date ?xsddate.
            bind( year(?xsddate) as ?year)
    filter (?year < 1900)

    }
}
}
group by ?year
order by ?year
```

Abbildung 12.7: sparql:metalEreignisseZeit

Kapitel 13

Notizbücher

Die Notizbücher sind unter

- https://gitlab.gwdg.de/dirk.wintergruen/netzwerkanalysen-und-semantische-datenmodellierung

archiviert und werden ausführlicher unter

- http://doi.org/21.11101/0000-0007-D3F9-2

erläutert.

Pythonnotizbücher

Kapitel 14

Indizes

OWL-Klassen, OWL-Eigenschaften und Named Graphs

Index

Kapitel 15

Bibliographie

Literatur

[1] *ABBYY Recognition Server 4*. URL: https://www.abbyy.com/en-us/support/regserv/4/pl/sr/ (besucht am 20. 07. 2018).

[2] *About the Newton Project*. URL: http://www.newtonproject.ox.ac.uk/about-us/newton-project (besucht am 18. 10. 2017).

[3] *ADS Search*. URL: https://ui.adsabs.harvard.edu/ (besucht am 08. 12. 2017).

[4] *Adsabs-Dev-Api: Developer API Service Description and Example Client Code*. SAO/NASA ADS, 10. Nov. 2017. URL: https://github.com/adsabs/adsabs-dev-api (besucht am 08. 12. 2017).

[5] Historische Commission bei der königl. Akademie der Wissenschaften. “Clavius, Christoph”. In: *Allgemeine Deutsche Biographie, Bd. 4*. 1. Allgemeine Deutsche Biographie. München/Leipzig: Duncker & Humblot, 1876, S. 298. URL: https://de.wikisource.org/wiki/ADB:Clavius,_Christoph.

[6] Réka Albert und Albert-László Barabási. “Statistical Mechanics of Complex Networks”. In: *Reviews of Modern Physics* 74.1 (30. Jan. 2002), S. 47–97. ISSN: 0034-6861, 1539-0756. DOI: 10.1103/RevModPhys.74.47.

[7] Jürgen Angele, Michael Kifer und Georg Lausen. “Ontologies in F-Logic”. In: *Handbook on Ontologies*. Hrsg. von Steffen Staab und Rudi Studer. International Handbooks on Information Systems. Springer Berlin Heidelberg, 2009, S. 45–70. ISBN: 978-3-540-70999-2 978-3-540-92673-3. DOI: 10.1007/978-3-540-92673-3_2. URL: http://link.springer.com/chapter/10.1007/978-3-540-92673-3_2 (besucht am 21. 11. 2016).

[8] Grigoris Antoniou und Frank van Harmelen. “Web Ontology Language: OWL”. In: *Handbook on Ontologies*. Hrsg. von Steffen Staab und Rudi Studer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, S. 91–110. ISBN: 978-3-540-92673-3. DOI: 10.1007/978-3-540-92673-3_4. URL: https://doi.org/10.1007/978-3-540-92673-3_4.

[9] *Apache Solr* -. URL: http://lucene.apache.org/solr/ (besucht am 30. 10. 2017).

[10] *Apache Solr vs Elasticsearch - the Feature Smackdown!* URL: http://solr-vs-elasticsearch.com/ (besucht am 30. 10. 2017).

[11] *Arboreal*. URL: http://archimedes.fas.harvard.edu/arboreal/ (besucht am 19. 10. 2017).

[12] *Architekturen des Wissens*. URL: /de/content/architekturen-des-wissens/ (besucht am 30. 10. 2017).

[13] *Art & Architecture Thesaurus (Getty Research Institute)*. URL: http://www.getty.edu/research/tools/vocabularies/aat/ (besucht am 27. 12. 2017).

[14] Albert-László Barabási. *Network Science by Albert-László Barabási*. 2017. URL: http://barabasi.com/networksciencebook/ (besucht am 03. 09. 2017).

[15] Albert-László Barabási und Réka Albert. "Emergence of Scaling in Random Networks". In: *Science* 286.5439 (15. Okt. 1999), S. 509–512. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.286.5439.509. pmid: 10521342.

[16] Juan A Barceló. *Computational Intelligence in Archaeology*. Hershey, PA [u.a.]: Information Science Reference, 2009. ISBN: 978-1-59904-489-7.

[17] A. Barrat und M. Weigt. "On the Properties of Small-World Network Models". In: *The European Physical Journal B - Condensed Matter and Complex Systems* 13.3 (1. Feb. 2000), S. 547–560. ISSN: 1434-6036. DOI: 10.1007/s100510050067.

[18] Chyssoula Bekiari, Martin Doerr und Patrick Le Boeuf, Hrsg. *FRBR - Object-Oriented Definition and Mapping to FRBRoo (Version 0.9 Draft)*. Jan. 2008.

[19] Elisa Bellotti, Luigi Guadalupi und Guido Conaldi. "Comparing Fields of Sciences: Multilevel Networks of Research Collaborations in Italian Academia". In: *Multilevel Network Analysis for the Social Sciences*. Methodos Series. Springer, Cham, 2016, S. 213–244. ISBN: 978-3-319-24518-8 978-3-319-24520-1. DOI: 10.1007/978-3-319-24520-1_9. URL: https://link.springer.com/chapter/10.1007/978-3-319-24520-1_9 (besucht am 20. 08. 2017).

[20] Elisa Bellotti, Luka Kronegger und Luigi Guadalupi. "The Evolution of Research Collaboration within and across Disciplines in Italian Academia". In: *Scientometrics* 109.2 (Nov. 2016), S. 783–811. ISSN: 0138-9130. DOI: 10.1007/s11192-016-2068-1.

[21] Luis M. A. Bettencourt und David I. Kaiser. "Formation of Scientific Fields as a Universal Topological Transition". In: (1. Apr. 2015). arXiv: 1504.00319 [physics]. URL: http://arxiv.org/abs/1504.00319 (besucht am 06. 11. 2015).

[22] Luis M. A. Bettencourt und David I. Kaiser. *General Critical Properties of the Dynamics of Scientific Discovery*. 1015864. 31. Mai 2011. URL: http://www.osti.gov/servlets/purl/1015864-VvU9MO/ (besucht am 06. 11. 2015).

[23] Luís M. A. Bettencourt, David I. Kaiser und Jasleen Kaur. "Scientific Discovery and Topological Transitions in Collaboration Networks". In: *Journal of Informetrics*. Science of Science: Conceptualizations and Models of Science 3.3 (Juli 2009), S. 210–221. ISSN: 1751-1577. DOI: 10.1016/j.joi.2009.03.001.

[24] Luis M. A. Bettencourt u. a. "Population Modeling of the Emergence and Development of Scientific Fields". In: *Scientometrics* 75.3 (1. Juni 2008), S. 495–518. ISSN: 0138-9130. DOI: 10.1007/s11192-007-1888-4.

[25] Luís M. A. Bettencourt u. a. “The Power of a Good Idea: Quantitative Modeling of the Spread of Ideas from Epidemiological Models”. In: *Physica A: Statistical Mechanics and its Applications* 364 (15. Mai 2006), S. 513–536. ISSN: 0378-4371. DOI: 10.1016/j.physa.2005.08.083.

[26] Deutsche Biographie. *Cholinus, Peter - Deutsche Biographie*. URL: https://www.deutsche-biographie.de/sfz8222.html (besucht am 20. 07. 2018).

[27] Steven Bird, Ewan Klein und Edward Loper. *Natural Language Processing with Python*. 2009. ISBN: 978-0-596-51649-9 0-596-51649-5.

[28] Tobias Blanke und Mark Hedges. “Scholarly Primitives: Building Institutional Infrastructure for Humanities e-Science”. In: *Special section: Recent advances in e-Science* 29.2 (1. Feb. 2013), S. 654–661. ISSN: 0167-739X. DOI: 10.1016/j.future.2011.06.006.

[29] *Blazegraph Download - Graph Database & Application Download*. URL: https://www.blazegraph.com/ (besucht am 10. 06. 2018).

[30] Alexander S. Blum, Roberto Lalli und Jürgen Renn. “The Renaissance of General Relativity: How and Why It Happened”. In: *Annalen der Physik* 528.5 (1. Mai 2016), S. 344–349. ISSN: 1521-3889. DOI: 10.1002/andp.201600105.

[31] Francis Bond und Ryan Foster. “Linking and Extending an Open Multilingual Wordnet”. In: *51st Annual Meeting of the Association for Computational Linguistics*. ACL-2013. Sofia, 2013, S. 1352–1362.

[32] Francis Bond und Kyonghee Paik. “A Survey of Wordnets and Their Licenses”. In: *Proceedings of the 6th Global WordNet Conference*. GWC 2012. Matsue, 2012, S. 64–71.

[33] Thomas Breuel. *Ocropy: Python-Based Tools for Document Analysis and OCR*. 19. Juli 2018. URL: https://github.com/tmbdev/ocropy (besucht am 20. 07. 2018).

[34] Gerhard Brewka. “The Logic of Inheritance in Frame Systems.” In: *Proceedings of the 10th International Joint Conference on Artificial Intelligence*. 10th International Joint Conference on Artificial Intelligence. 1987, S. 483–488.

[35] William J Browne, Harvey Goldstein und Jon Rasbash. “Multiple Membership Multiple Classification (MMMC) Models”. In: *Statistical Modelling* 1.2 (1. Juli 2001), S. 103–124. ISSN: 1471-082X. DOI: 10.1177/1471082X0100100202.

[36] Natasa Bulatovic u. a. *Digital Scrapbook: Can We Enable Interlinked and Recursive Knowledge Equilibrium? :* Berlin: Max-Planck-Institut für Wissenschaftsgeschichte, 2015. URL: http://www.mpiwg-berlin.mpg.de/Preprints/P474.PDF.

[37] “Bulletin on General Relativity and Gravitation”. In: *Bulletin on general relativity and gravitation* (2010). ISSN: 1662-5390.

[38] Jeremy J. Carroll u. a. “Named Graphs, Provenance and Trust”. In: *Proceedings of the 14th International Conference on World Wide Web*. WWW ’05. New York, NY, USA: ACM, 2005, S. 613–622. ISBN: 1-59593-046-9. DOI: 10.1145/1060745.1060835.

[39] Andy Casey. *Ads: Python Tool for ADS*. 15. Okt. 2017. URL: https://github.com/andycasey/ads (besucht am 08. 12. 2017).

[40] *CDLI - Cuneiform Digital Library Initiative*. URL: https://cdli.ucla.edu/ (besucht am 18. 10. 2017).

[41] Chaomei Chen. "CiteSpace II: Detecting and Visualizing Emerging Trends and Transient Patterns in Scientific Literature". In: *Journal of the American Society for Information Science and Technology* 57.3 (1. Feb. 2006), S. 359–377. ISSN: 1532-2890. DOI: 10.1002/asi.20317.

[42] Chaomei Chen. *How to Use CiteSpace*. Leanpub, 23. Nov. 2015. URL: https://leanpub.com/howtousecitespace (besucht am 16. 01. 2016).

[43] Chaomei Chen, Fidelia Ibekwe-SanJuan und Jianhua Hou. "The Structure and Dynamics of Cocitation Clusters: A Multiple-Perspective Cocitation Analysis". In: *Journal of the American Society for Information Science and Technology* 61.7 (1. Juli 2010), S. 1386–1409. ISSN: 1532-2890. DOI: 10.1002/asi.21309.

[44] Chaomei Chen u. a. "Towards an Explanatory and Computational Theory of Scientific Discovery". In: (8. Apr. 2009). arXiv: 0904.1439 [cs]. URL: http://arxiv.org/abs/0904.1439 (besucht am 29. 05. 2017).

[45] Wai-Ki Ching und Michael K. Ng. *Markov Chains: Models, Algorithms and Applications*. International Series in Operations Research & Management Science. Springer US, 2006. ISBN: 978-1-4419-3986-9. URL: //www.springer.com/de/book/9781441939869 (besucht am 30. 07. 2018).

[46] Aaron Clauset, Cristopher Moore und M. E. J. Newman. "Hierarchical Structure and the Prediction of Missing Links in Networks". In: *Nature* 453.7191 (1. Mai 2008), S. 98–101. ISSN: 0028-0836. DOI: 10.1038/nature06830.

[47] Aaron Clauset, M. E. J. Newman und Cristopher Moore. "Finding Community Structure in Very Large Networks". In: *Physical Review E* 70.6 (6. Dez. 2004). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.70.066111. arXiv: cond-mat/0408187.

[48] *CVS - Open Source Version Control*. URL: http://www.nongnu.org/cvs/ (besucht am 18. 10. 2017).

[49] *Cython: C-Extensions for Python*. URL: http://cython.org/ (besucht am 09. 06. 2018).

[50] *Cytoscape: An Open Source Platform for Complex Network Analysis and Visualization*. URL: http://www.cytoscape.org/ (besucht am 09. 06. 2018).

[51] *(D-6) Atlas of Innovations (Key Topic Innovation) | Topoi*. 2016. URL: https://www.topoi.org/group/d-6/ (besucht am 22. 11. 2016).

[52] Peter Damerow und Wolfgang Lefèvre. *Wissenssysteme im geschichtlichen Wandel :* Berlin: Max-Planck-Institut für Wissenschaftsgeschichte, 1994. URL: http://www.mpiwg-berlin.mpg.de/Preprints/P5.PDF.

[53] *Darwin Correspondence Project*. URL: https://www.darwinproject.ac.uk/ (besucht am 18. 10. 2017).

[54] Lorraine Daston und Michael Stolleis. *Natural Law And Laws of Nature in Early Modern Europe: Jurisprudence, Theology, Moral And Natural Philosophy*. Ashgate Publishing Limited, 22. Dez. 2008. ISBN: 0-7546-5761-2.

[55] *Database: Blazegraph High Performance Graph Database*. Blazegraph, 8. Juni 2018. URL: https://github.com/blazegraph/database (besucht am 10. 06. 2018).

[56] *Dataverse-Client-Python: Python Library for Writing Clients That Use APIs from Dataverse*. Institute for Quantitative Social Science, 30. Mai 2018. URL: https://github.com/IQSS/dataverse-client-python (besucht am 09. 06. 2018).

[57] *De Sphaera | The Sphere*. URL: https://sphaera.mpiwg-berlin.mpg.de/ (besucht am 13. 04. 2019).

[58] Gerard de Melo und Gerhard Weikum. "Untangling the Cross-Lingual Link Structure of Wikipedia". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Uppsala, Sweden: Association for Computational Linguistics, Juli 2010, S. 844–853. ISBN: 978-1-932432-66-4. URL: http://www.aclweb.org/anthology/P10-1087.

[59] Meghnad Desai u. a. "Measuring the Technology Achievement of Nations and the Capacity to Participate in the Network Age". In: *Journal of Human Development* 3.1 (1. Feb. 2002), S. 95–122. ISSN: 1464-9888. DOI: 10.1080/14649880120105399.

[60] *Deutsche Nationalbibliothek - Datendienst*. URL: http://www.dnb.de/DE/Service/DigitaleDienste/Datendienst/datendienst_node.html (besucht am 10. 06. 2018).

[61] *Deutsche Nationalbibliothek - GND*. URL: http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html (besucht am 10. 06. 2018).

[62] *Django CMS - Enterprise Content Management with Django - Django CMS*. URL: https://www.django-cms.org/en/ (besucht am 09. 06. 2018).

[63] Quinn Dombrowski. "What Ever Happened to Project Bamboo?" In: *Literary and Linguistic Computing* 29.3 (1. Sep. 2014), S. 326–339. ISSN: 0268-1145. DOI: 10.1093/llc/fqu026.

[64] D-PLACE-Consortium. *D-PLACE*. 2016. URL: https://d-place.org/ (besucht am 22. 11. 2016).

[65] *Dropbox - Dropbox Business*. URL: https://www.dropbox.com/business (besucht am 09. 06. 2018).

[66] *Drupal - Open Source CMS*. 5. Apr. 2018. URL: https://www.drupal.org/home (besucht am 09. 06. 2018).

[67] Marten Düring. *Historical Network Research*. 7. Jan. 2013. URL: http://historicalnetworkresearch.org/ (besucht am 12. 07. 2016).

[68] Marten Düring u. a., Hrsg. *Handbuch Historische Netzwerkforschung: Grundlagen und Anwendungen*. Berlin [u.a.]: LIT-Verl., 2016. ISBN: 978-3-643-11705-2.

[69] *ECHO*. URL: http://echo.mpiwg-berlin.mpg.de/home (besucht am 18. 10. 2017).

[70] D. Ediger u. a. “Massive Social Network Analysis: Mining Twitter for Social Good”. In: *2010 39th International Conference on Parallel Processing*. 2010 39th International Conference on Parallel Processing. Sep. 2010, S. 583–593. DOI: 10.1109/ICPP.2010.66.

[71] Jean Eisenstaedt. “The Low Water Mark of General Relativity, 1925-1955”. In: *Einstein and the History of General Relativity*. Hrsg. von D. Howard und John Stachel. Birkhäuser, 1989, S. 1–277.

[72] Scott Emmons u. a. “Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale”. In: *PLoS ONE* 11.7 (8. Juli 2016). ISSN: 1932-6203. DOI: 10.1371/journal.pone.0159161. pmid: 27391786.

[73] P. Erdős und A Rényi. “On the Evolution of Random Graphs”. In: *PUBLICATION OF THE MATHEMATICAL INSTITUTE OF THE HUNGARIAN ACADEMY OF SCIENCES*. 1960, S. 17–61.

[74] *Europeana Collections*. URL: https://www.europeana.eu/portal/?locale=en (besucht am 18. 10. 2017).

[75] Stefan Evert u. a. “Understanding and Explaining Delta Measures for Authorship Attribution”. In: *Digital Scholarship in the Humanities* 32 (suppl_2 1. Dez. 2017), S. ii4–ii16. ISSN: 2055-7671. DOI: 10.1093/llc/fqx023.

[76] *Extended Open Multilingual Wordnet*. URL: http://compling.hss.ntu.edu.sg/omw/summx.html (besucht am 11. 08. 2017).

[77] *FactForge*. URL: http://factforge.net/sparql (besucht am 09. 08. 2017).

[78] Heiner Fangerau. “Evolution of Knowledge from a Network Perspective: Recognition as a Selective Factor in the History of Science”. In: (2013), S. 11–32.

[79] Heiner Fangerau. *Spinning the scientific web: Jacques Loeb (1859-1924) und sein Programm einer internationalen biomedizinischen Grundlagenforschung*. Akademie-Verlag, 2010. 284 S. ISBN: 978-3-05-004528-3.

[80] *Fedora Repository: Flexible. Modular. Open-Source.* URL: https://duraspace.org/fedora/ (besucht am 09. 06. 2018).

[81] Chris Fehily. *SQL: Visual QuickStart Guide*. Peachpit Press, 19. Juli 2002. 424 S. ISBN: 978-0-321-11803-5. URL: http://proquest.tech.safaribooksonline.de/0321118030 (besucht am 30. 10. 2017).

[82] Giorgos Flouris u. a. “Coloring RDF Triples to Capture Provenance”. In: *The Semantic Web - ISWC 2009*. Hrsg. von Abraham Bernstein u. a. Bearb. von David Hutchison u. a. Bd. 5823. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, S. 196–212. ISBN: 978-3-642-04929-3 978-3-642-04930-9. URL: http://rd.springer.com/chapter/10.1007/978-3-642-04930-9_13 (besucht am 02. 09. 2013).

[83] Robert William Fogel. “The Limits of Quantitative Methods in History”. In: *The American Historical Review* 80.2 (1975), S. 329–350. ISSN: 0002-8762. DOI: 10.2307/1850498. JSTOR: 1850498.

[84] *Functional Overview | CIDOC CRM*. URL: http://www.cidoc-crm.org/functional-units (besucht am 25. 11. 2017).

[85] *Galileo Galilei, Ms. Gal. 72*. URL: https://www.mpiwg-berlin.mpg.de/Galileo_Prototype/INDEX.HTM (besucht am 18. 10. 2017).

[86] Peter Gärdenfors. “The Dynamics of Belief Systems: Foundations vs. Coherence Theories”. In: *Revue internationale de philosophie* 44 (1992), S. 24–46.

[87] *Gephi - The Open Graph Viz Platform*. URL: https://gephi.org/ (besucht am 09. 06. 2018).

[88] *Getty Vocabularies as LOD (Getty Research Institute)*. URL: http://www.getty.edu/research/tools/vocabularies/lod/index.html (besucht am 12. 08. 2017).

[89] *Git*. URL: https://git-scm.com/ (besucht am 21. 04. 2019).

[90] Paul Gooding, Melissa Terras und Claire Warwick. “The Myth of the New: Mass Digitization, Distant Reading, and the Future of the Book”. In: *Literary and Linguistic Computing* 28.4 (12. Jan. 2013), S. 629–639. ISSN: 0268-1145, 1477-4615. DOI: 10.1093/llc/fqt051.

[91] *Google Drive – Cloud-Speicherplatz und Datensicherung für Fotos, Dokumente und mehr*. URL: www.google.com/intl/de_ALL/drive/ (besucht am 09. 06. 2018).

[92] Günther Görz. “”Generics and Defaults”. Zum technischen Umgang mit Begriffssystemen, Standardannahmen und Ausnahmen.” In: *Methodisches Denken im Kontext. Festschrift für Christian Thiel zum 70. Geburtstag.* Hrsg. von Volker Peckhaus und Peter Bernhard. Paderborn: mentis, 2007, S. 383–401.

[93] Günther Görz, Hrsg. *Handbuch Der Künstlichen Intelligenz*. 5., überarb. und aktualisierte Aufl. München: Oldenbourg, 2014. ISBN: 978-3-486-71307-7.

[94] Günther Görz, Bernhard Schiemann und Martin Oischinger. “An Implementation of the CIDOC Conceptual Reference Model (4.2.4) in OWL-DL”. In: *Proceedings CIDOC 2008 — The Digital Curation of Cultural Heritage. (CIDOC 2008 — The Digital Curation of Cultural Heritage. Athen, Benaki Museum 15.-18.09.2008)*. Hrsg. von Angelos Delivorrias. Athens: ICOM CIDOC, 2008, S. 1–14.

[95] C. Graham. *Markov Chains Analytic and Monte Carlo Computations*. Chichester, West Sussex: Wiley, 2014. URL: http://proquest.tech.safaribooksonline.de/?uiCode=planck&xmlId=9781118882696.

[96] *Graph-Tool: Efficent Network Analysis with Python*. URL: https://graph-tool.skewed.de/ (besucht am 09. 06. 2018).

[97] Gerd Graßhoff. “The Discovery of the Urea Cycle: Computer Models of Scientific Discovery”. In: *Computer Simulations in Science and Technology Studies*. Hrsg. von Petra Ahrweiler und Nigel Gilbert. Springer Berlin Heidelberg, 1. Jan. 1998, S. 71–90. ISBN: 978-3-642-63521-2. URL: http://dx.doi.org/10.1007/978-3-642-58270-7_5.

[98] Gerd Graßhoff. “The Historical Basis of Scientific Discovery”. In: *Behavioral and Brain Sciences* 17.03 (1994), S. 545–546. DOI: 10.1017/S0140525X00035846.

[99] Gerd Graßhoff und Michael May. “From Historical Case Studies to Systematic Methods of Discovery”. In: *Working notes: AAAI Spring Symposium on Systematic Methods of Scientific Discovery* (1995).

[100] Gerd Graßhoff und Michael May. “Hans Krebs’ and Kurt Hanseleit ’s Laboratory Notebooks and Their Discovery of the Urea Cycle - Recontructed with Computer Models”. In: *Reworking the Bench: Research Notebooks in the History of Science*. Unter Mitarb. von Frederic L. Holmes, Jürgen Renn und Hans-Jorg Rheinberger. Klüwer, 2003, S. 269–294.

[101] Jonathan Gross, Jay Yellen und Ping Zhang. *Handbook of Graph Theory, Second Edition*. 2. Aufl. Chapman and Hall/CRC, 2013. 1630 S. ISBN: 978-1-4398-8018-0. URL: http://proquest.tech.safaribooksonline.de/9781439880197 (besucht am 23. 11. 2016).

[102] Martin Grötschel. “Schnelle Rundreisen: Das Travelling-Salesman-Problem”. In: *Kombinatorische Optimierung erleben*. Hrsg. von Prof Dr Stephan Hußmann und Dr Brigitte Lutz-Westphal. Vieweg+Teubner, 2007, S. 95–129. ISBN: 978-3-528-03216-6 978-3-8348-9120-4. DOI: 10.1007/978-3-8348-9120-4_4. URL: http://link.springer.com/chapter/10.1007/978-3-8348-9120-4_4 (besucht am 08. 11. 2016).

[103] ASU Digital Innovation Group. *Tethne 0.8 Documentation — Tethne 0.8 Documentation*. 2015. URL: http://diging.github.io/tethne/ (besucht am 06. 06. 2017).

[104] CIDOC CRM Special Interest Group. *CIDOC CRM*. 2015. URL: http://www.cidoc-crm.org/Version/version-6.2.1 (besucht am 23. 11. 2016).

[105] Thomas R. Gruber. “Toward Principles for the Design of Ontologies Used for Knowledge Sharing?” In: *International Journal of Human-Computer Studies* 43.5 (1. Nov. 1995), S. 907–928. ISSN: 1071-5819. DOI: 10.1006/ijhc.1995.1081.

[106] Nicola Guarino, Daniel Oberle und Steffen Staab. “What Is an Ontology?” In: *Handbook on Ontologies*. International Handbooks on Information Systems. Springer, Berlin, Heidelberg, 2009, S. 1–17. ISBN: 978-3-540-70999-2 978-3-540-92673-3. DOI: 10.1007/978-3-540-92673-3_0. URL: https://link.springer.com/chapter/10.1007/978-3-540-92673-3_0 (besucht am 30. 10. 2017).

[107] Heinz-Peter Gumm, Manfred Sommer und Wofgang Hesse. *Einführung in die Informatik*. 5. Ausgabe. OCLC: 866056685. München: Oldenbourg, 2002. ISBN: 978-3-486-71995-6.

[108] Mark Handcock u. a. “Statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data”. In: *Journal of Statistical Software, Articles* 24.1 (2008), S. 1–11. ISSN: 1548-7660. DOI: 10.18637/jss.v024.i01.

[109] *HermiT Reasoner: Home*. URL: http://www.hermit-reasoner.com/ (besucht am 23. 04. 2019).

[110] Pascal Hitzler, Markus Krötzsch und Sebastian Rudolph. *Foundations of Semantic Web Technologies*. CRC Press, Aug. 2009. ISBN: 978-1-4200-9050-5.

[111] *HOBBIT — Agile Knowledge Engineering and Semantic Web (AKSW)*. URL: http://aksw.org/Projects/HOBBIT.html (besucht am 15. 08. 2017).

[112] Jake M. Hofman und Chris H. Wiggins. "Bayesian Approach to Network Modularity". In: *Physical Review Letters* 100.25 (23. Juni 2008), S. 258701. DOI: 10.1103/PhysRevLett.100.258701.

[113] *http://wordnet.rkbexplorer.com/sparql/*. URL: http://wordnet.rkbexplorer.com/sparql/ (besucht am 09. 08. 2017).

[114] David Hunter u. a. "Ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks". In: *Journal of Statistical Software, Articles* 24.3 (2008), S. 1–29. ISSN: 1548-7660. DOI: 10.18637/jss.v024.i03.

[115] Malcolm D. Hyman. *Semantic Networks: A Tool for Investigating Conceptual Change and Knowledge Transfer in the History of Science*. Bd. Übersetzung und Transformation Hrsg. v. Böhme, Hartmut / Rapp, Christof / Rösler, Wolfgang. Transformationen Der Antike. Berlin: De Gruyter, 2007.

[116] *iCloud*. URL: https://www.apple.com/de/icloud/ (besucht am 09. 06. 2018).

[117] *ICS - CRMinf: The Argumentation Model*. URL: http://www.ics.forth.gr/isl/index_main.php?l=e&c=713 (besucht am 30. 10. 2017).

[118] *Igraph – Network Analysis Software*. URL: http://igraph.org/ (besucht am 05. 09. 2017).

[119] *Index of /Wikidatawiki/Entities/*. URL: https://dumps.wikimedia.org/wikidatawiki/entities/ (besucht am 10. 06. 2018).

[120] Matthew O. Jackson. *Social and Economic Networks*. Princeton and N.J: Princeton University Press, 2008. ISBN: 978-0-691-13440-6.

[121] Dorothea Jansen. *Einführung in die Netzwerkanalyse - Grundlagen, Methoden, Forschungsbeispiele*. 2003. URL: http://link.springer.com/book/10.1007/978-3-663-09875-1 (besucht am 15. 03. 2015).

[122] Dorothea Jansen. "Netzwerkanalyse, soziale Strukturen und soziales Kapital". In: *Einführung in die Netzwerkanalyse*. VS Verlag für Sozialwissenschaften, 2003, S. 11–36. ISBN: 978-3-8100-3149-5 978-3-663-09875-1. URL: http://link.springer.com/chapter/10.1007/978-3-663-09875-1_1 (besucht am 15. 03. 2015).

[123] *Jenkins*. URL: https://jenkins.io/index.html (besucht am 10. 06. 2018).

[124] Philip N Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference and Consciousness*. Cambridge [Cambridgeshire]: Cambridge University Press, 1983. 513 S. ISBN: 0-521-24123-5.

[125] Steven E Jones. *Roberto Busa, S.J., and the Emergence of Humanities Computing: The Priest and the Punched Cards*. OCLC: 921310798. 2016. ISBN: 978-1-138-18677-4.

[126] Rudolf Juchhoff. “Cholinus. Maternus”. In: *Neue Deutsche Biographie, Band 3*. 1957, 213 [Online Version]. URL: https://www.deutsche-biographie.de/pnd119640619.html#ndbcontent.

[127] Project Jupyter. *Project Jupyter*. 2016. URL: http://www.jupyter.org (besucht am 25. 11. 2016).

[128] Andrew Keen. *The Internet Is Not the Answer*. OCLC: 929606251. 2015. ISBN: 978-1-78239-343-6.

[129] Daniel Keim. *Mastering the Information Age: Solving Problems with Visual Analytics*. OCLC: 835305616. Goslar: Eurographics Association, 2010. ISBN: 978-3-905673-77-7.

[130] Florian Kerschbaumer und Marten Düring. “Quantifizierung und Visualisierung. Anknüpfungspunkte in den Geschichtswissenschaften”. In: (2016). URL: http://orbilu.uni.lu/handle/10993/31307 (besucht am 11. 05. 2018).

[131] Jon Kleinberg. “Bursty and Hierarchical Structure in Streams”. In: *Data Mining and Knowledge Discovery* 7.4 (1. Okt. 2003), S. 373–397. ISSN: 1384-5810, 1573-756X. DOI: 10.1023/A:1024940629314.

[132] Achim Klenke. *Wahrscheinlichkeitstheorie*. Hrsg. von SpringerLink (Online service). 3. Aufl. 2013. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. ISBN: 978-3-642-36018-3. URL: http://dx.doi.org/10.1007/978-3-642-36018-3?nosfx=y.

[133] Donald Ervin Knuth. *Literate Programming*. OCLC: 24906757. Stanford, Calif.: Center for the Study of Language and Information, 1992. ISBN: 978-0-937073-80-3 978-0-937073-81-0.

[134] Jürgen Kocka. “Theories and Quantification in History”. In: *Social Science History* 8.2 (1984), S. 169–178. ISSN: 0145-5532. DOI: 10.2307/1170992. JSTOR: 1170992.

[135] Florian Kräutli und Matteo Valleriani. “CorpusTracer: A CIDOC Database for Tracing Knowledge Networks”. In: *Digital Scholarship in the Humanities* (2017). DOI: 10.1093/llc/fqx047.

[136] Roberto Lalli. *Building the General Relativity and Gravitation Community during the Cold War*. Berlin: Springer, 2017. ISBN: 978-3-319-54653-7 3-319-54653-8.

[137] Manfred Laubichler und Jürgen Renn. “Extended Evolution: A Conceptual Framework for Integrating Regulatory Networks and Niche Construction”. In: *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 324.7 (2015), S. 1–13. ISSN: 1552-5007.

[138] Emmanuel Lazega und Tom A.B. Snijders, Hrsg. *Multilevel Network Analysis for the Social Sciences*. Cham: Springer International Publishing, 2016. ISBN: 978-3-319-24518-8 978-3-319-24520-1. URL: http://link.springer.com/10.1007/978-3-319-24520-1 (besucht am 09. 04. 2016).

[139] Oona Leganovic u. a. *Anforderungen von Geisteswissenschaftlern an einen digitalen Forschungsprozess*. 1. Feb. 2015.

[140] Claire Lemercier. “Formale Methoden der Netzwerkanalyse in den Geschichtswissenschaften: Warum und Wie?” In: *Historische Netzwerkanalyse*. Hrsg. von Albert Müller und Wolfgang Neurath. Innsbruck: Studien Verlag, 2012.

[141] *Lemon - Lexicon Model for Ontologies*. URL: http://lemon-model.net/ (besucht am 14. 12. 2017).

[142] Richard Lenz. “Evolutionäre Informationssysteme”. Habilitationsschrift. Marburg, 2005.

[143] *Linked Data Platform 1.0*. URL: https://www.w3.org/TR/ldp/ (besucht am 25. 11. 2017).

[144] *Linked Open Vocabularies (LOV)*. URL: http://lov.okfn.org/dataset/lov/vocabs/rdfg (besucht am 26. 12. 2017).

[145] Yabing Liu u. a. “Analyzing Facebook Privacy Settings: User Expectations vs. Reality”. In: *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*. IMC ’11. Berlin, Germany: ACM, 2011, S. 61–70. ISBN: 978-1-4503-1013-0. DOI: 10.1145/2068816.2068823.

[146] *Mapping the Republic of Letters*. URL: http://republicofletters.stanford.edu/ (besucht am 12. 07. 2016).

[147] *Marian Dörk*. URL: http://mariandoerk.de/ (besucht am 30. 07. 2018).

[148] Michael Matuschek und Iryna Gurevych. “Dijkstra-WSA: A Graph-Based Approach to Word Sense Alignment”. In: *Transactions of the Association for Computational Linguistics (TACL)* 1 (Mai 2013), S. 151–164.

[149] Andreas Meier und Michael Kaufmann. *SQL- & NoSQL-Datenbanken*. eXamen.Press. Berlin, Heidelberg: Springer Berlin Heidelberg, 2016. ISBN: 978-3-662-47663-5 978-3-662-47664-2. DOI: 10.1007/978-3-662-47664-2. URL: http://link.springer.com/10.1007/978-3-662-47664-2 (besucht am 30. 10. 2017).

[150] Gerard de Melo und Gerhard Weikum. “Towards a Universal Wordnet by Learning from Combined Evidence”. In: *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*. Hong Kong, China: ACM, 2009, S. 513–522. ISBN: 978-1-60558-512-3. DOI: http://doi.acm.org/10.1145/1645953.1646020.

[151] *Mercurial SCM*. URL: https://www.mercurial-scm.org/ (besucht am 21. 04. 2019).

[152] *Metaphacts/Wikidata-Qald-7 - Docker Hub*. URL: https://hub.docker.com/r/metaphacts/wikidata-qald-7/ (besucht am 15. 08. 2017).

[153] George A. Miller. “WordNet: A Lexical Database for English”. In: *Commun. ACM* 38.11 (Nov. 1995), S. 39–41. ISSN: 0001-0782. DOI: 10.1145/219717.219748.

[154] Marvin Minsky. “A Framework for Representing Knowledge”. In: *Readings on Knowledge Representation*. Unter Mitarb. von R. J. Brachmann und H. J. Levesque. http://web.media.mit.edu/ minsky/papers/Frames/frames.html, 1975.

[155] Hans-Joachim Mittag. *Statistik: Eine Einführung mit interaktiven Elementen*. 3. Aufl. Springer-Lehrbuch. Springer Spektrum, 2014. ISBN: 978-3-642-54387-6. URL: //www.springer.com/us/book/9783642543876 (besucht am 30. 07. 2018).

[156] ModellingSD. *Scholarly_Domain: Modelling the Scholarly Domain*. 12. Apr. 2016. URL: https://github.com/ModellingSD/Scholarly_Domain (besucht am 15. 10. 2017).

[157] Franco Moretti. *Distant Reading*. 1. publ. London [u.a.]: Verso, 2013. ISBN: 978-1-78168-084-1.

[158] *Natural Language Toolkit — NLTK 3.2.4 Documentation*. URL: http://www.nltk.org/ (besucht am 11. 08. 2017).

[159] *Neo4j Open Source Nosql Graph Database »*. URL: http://neo4j.org/.

[160] *NetworkX — NetworkX*. URL: https://networkx.github.io/ (besucht am 05. 09. 2017).

[161] M. E. J. Newman. "Finding Community Structure in Networks Using the Eigenvectors of Matrices". In: *Physical Review E* 74.3 (11. Sep. 2006). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.74.036104. arXiv: physics/0605087.

[162] M. E. J. Newman. "The Structure and Function of Complex Networks". In: *SIAM Review* 45.2 (Jan. 2003), S. 167–256. ISSN: 0036-1445, 1095-7200. DOI: 10.1137/S003614450342480. arXiv: cond-mat/0303516.

[163] M. E. J. Newman und M. Girvan. "Finding and Evaluating Community Structure in Networks". In: *Physical Review E* 69.2 (26. Feb. 2004). ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.69.026113. arXiv: cond-mat/0308217.

[164] Mark E. J. Newman, Steven H. Strogatz und Duncan J. Watts. "Random Graphs with Arbitrary Degree Distributions and Their Applications". In: *Physical Review E* 64.2 (24. Juli 2001), S. 026118. DOI: 10.1103/PhysRevE.64.026118.

[165] Mark E. J. Newman und Duncan J. Watts. "Renormalization Group Analysis of the Small-World Network Model". In: *Physics Letters A* 263.4–6 (6. Dez. 1999), S. 341–346. ISSN: 0375-9601. DOI: 10.1016/S0375-9601(99)00757-4.

[166] Nextcloud. *The Most Popular Self-Hosted File Share and Collaboration Platform*. URL: http://nextcloud.com (besucht am 09. 06. 2018).

[167] Andrey Nikishaev. *Feature Extraction and Similar Image Search with OpenCV for Newbies*. URL: https://medium.com/machine-learning-world/feature-extraction-and-similar-image-search-with-opencv-for-newbies-3c59796bf774 (besucht am 19. 07. 2018).

[168] Hans J. Nissen, Peter Damerow und Robert K. Englund. *Archaic Bookkeeping: Early Writing and Techniques of Economic Administration in the Ancient Near East*. University of Chicago Press, Jan. 1993. 196 S. ISBN: 978-0-226-58659-5.

[169] *Nltk.Chunk Package — NLTK 3.2.4 Documentation*. URL: http://www.nltk.org/api/nltk.chunk.html?highlight=regex#nltk.chunk.regexp.RegexpParser (besucht am 11. 08. 2017).

[170] *Notebook Basics—Wolfram Language Documentation*. URL: http://reference.wolfram.com/language/guide/NotebookBasics.html (besucht am 06. 05. 2018).

[171] Marc Nunkesser und Daniel Sawitzki. “Blockmodels”. In: *Network Analysis*. Hrsg. von Ulrik Brandes und Thomas Erlebach. Lecture Notes in Computer Science 3418. Springer Berlin Heidelberg, 2005, S. 253–292. ISBN: 978-3-540-24979-5 978-3-540-31955-9. DOI: 10.1007/978-3-540-31955-9_10. URL: http://link.springer.com/chapter/10.1007/978-3-540-31955-9_10 (besucht am 15. 06. 2016).

[172] *Ocrocis*. URL: http://cistern.cis.lmu.de/ocrocis/ (besucht am 20. 07. 2018).

[173] *Open Source Search & Analytics · Elasticsearch*. URL: http://www.elastic.co (besucht am 30. 10. 2017).

[174] *OpenCV Library*. URL: https://opencv.org/ (besucht am 19. 07. 2018).

[175] *OpenID Connect | OpenID*. URL: http://openid.net/connect/ (besucht am 09. 06. 2018).

[176] Peter Oram. “WordNet: An Electronic Lexical Database. Christiane Fellbaum (Ed.). Cambridge, MA: MIT Press, 1998. Pp. 423. -”. In: *Applied Psycholinguistics* 22.1 (2001), S. 131–134.

[177] Jacco van Ossenbruggen. *Wordnet-3.0-Rdf: The Linked Open Dataset Described at http://datahub.io/datas wordnet, and the Tools Used to Create It*. 20. Juni 2017. URL: https://github.com/jrvosse/wordnet-3.0-rdf (besucht am 09. 08. 2017).

[178] *OWL : FaCT++*. URL: http://owl.man.ac.uk/factplusplus/ (besucht am 23. 04. 2019).

[179] *ownCloud - The Last Cloud Collaboration Platform You’ll Ever Need.* 27. Nov. 2017. URL: https://owncloud.org/ (besucht am 09. 06. 2018).

[180] *Oxygen XML Editor*. URL: https://www.oxygenxml.com/ (besucht am 01. 05. 2019).

[181] John F. Padgett und Christopher K. Ansell. “Robust Action and the Rise of the Medici, 1400-1434”. In: *American Journal of Sociology* 98.6 (1. Mai 1993), S. 1259–1319. ISSN: 0002-9602. JSTOR: 2781822.

[182] Eli Pariser und Ursula Held. *Filter Bubble wie wir im Internet entmündigt werden*. München: Hanser, 2017. ISBN: 978-3-446-45352-4 3-446-45352-0.

[183] *Pellet Is an OWL 2 Reasoner in Java; Open Source (AGPL) and Commercially Licensed, Commercial Support Available.: Stardog-Union/Pellet*. Stardog Union, 19. Apr. 2019. URL: https://github.com/stardog-union/pellet (besucht am 23. 04. 2019).

[184] *Perseus Digital Library*. URL: http://www.perseus.tufts.edu/hopper/ (besucht am 18. 10. 2017).

[185] Johann Pfanzagl. *Elementare Wahrscheinlichkeitsrechnung*. 2., überarbeitete und erweiterte Auflage, im Original erschienen 1991. Berlin ; New York: De Gruyter, 1991. ISBN: 978-3-11-013384-4. URL: http://www.reference-global.com/doi/book/10.1515/9783110869217.

[186] Karl R. Popper. *Logik Der Forschung*. Hrsg. von Herbert Keuth. 11. Aufl. Tübingen: Mohr Siebeck, 2005. xiii+601. ISBN: 978-3-16-148111-6.

[187] Johannes Preiser-Kapeller. “Calculating Byzantium?” In: *Social network analysis and complexity sciences as tools for the exploration of medieval social dynamics. Austrian Academy of Sciences Report* (2010). URL: http://www.oeaw.ac.at/byzanz/repository/Preiser_WorkingPapers_Calculating_I.pdf (besucht am 09. 05. 2016).

[188] Johannes Preiser-Kapeller. “Complex Historical Dynamics of Crisis: The Case of Byzantium”. In: *Krise Und Transformation*. Verlag der Österr. Akad. der Wiss., 2012, S. 69–127. URL: http://historicalnetworkresearch.org/wp-content/uploads/2013/01/Preiser_WorkingPapersIV_ComplexCrisis.pdf (besucht am 09. 05. 2016).

[189] Johannes Preiser-Kapeller. “Harbours and Maritime Mobility: Networks and Entanglements”. In: *Harbours and Maritime Networks as Complex Adaptive Systems*. 2015, S. 119–140.

[190] Johannes Preiser-Kapeller. “Harbours and Maritime Networks as Complex Adaptive Systems - a Thematic Introduction”. In: *Harbours and Maritime Networks as Complex Adaptive Systems*. 2015, S. 1–24.

[191] Johannes Preiser-Kapeller und Falko Daim. *Harbours and Maritime Networks as Complex Adaptive Systems*. Bd. 23. Römisch-Germanisches Zentralmuseum - Tagungen. Mainz, 2015.

[192] *Protégé*. URL: https://protege.stanford.edu/ (besucht am 15. 10. 2017).

[193] *Pulotu*. 2016. URL: https://pulotu.econ.mpg.de/ (besucht am 22. 11. 2016).

[194] *RDF 1.1 TriG*. URL: https://www.w3.org/TR/2014/REC-trig-20140225/ (besucht am 27. 11. 2017).

[195] *RDF Schema 1.1*. URL: https://www.w3.org/TR/rdf-schema/ (besucht am 27. 11. 2017).

[196] *RDF - Semantic Web Standards*. URL: https://www.w3.org/RDF/ (besucht am 30. 10. 2017).

[197] *React - A JavaScript Library for Building User Interfaces*. URL: https://reactjs.org/index.html (besucht am 10. 06. 2018).

[198] Wolfgang Reinhard. *Freunde und Kreaturen: ”Verflechtung” als Konzept zur Erforschung historischer Führungsgruppen: Römische Oligarchie um 1600*. Bd. 14. Schriften der Philosophischen Fachbereiche der Universität Augsburg. München: Vögel, 1979. ISBN: 978-3-920896-51-9.

[199] *RELAX NG Home Page*. URL: http://relaxng.org/ (besucht am 30. 10. 2017).

[200] Jürgen Renn. “Beyond Editions - Historical Sources in the Digital Age”. In: *Internationalität und Interdisziplinarität der Editionswissenschaft* (2014). ISSN: 9783110367317. DOI: 10.1515/9783110367317.17.

[201] Jürgen Renn. “Historical Epistemology and Interdisciplinarity”. In: *Physics, Philosophy, and the Scientific Community*. Boston Studies in the Philosophy of Science. Springer, Dordrecht, 1995, S. 241–251. ISBN: 978-90-481-4436-5 978-94-017-2658-0. DOI: 10.1007/978-94-017-2658-0_14. URL: https://link.springer.com/chapter/10.1007/978-94-017-2658-0_14 (besucht am 18. 10. 2017).

[202] Jürgen Renn. “Mentale Modelle in der Geschichte des Wissens: Auf dem Wege zu einer Paläontologie des mechanischen Denkens”. In: *Dahlemer Archivgespräche*. Unter Mitarb. von E. Henning. 6: Archiv zur Geschichte der Max-Planck-Gesellschaft, 2000.

[203] Jürgen Renn und Peter Damerow. “Mentale Modelle als kognitive Instrumente der Transformation von technischem Wissen”. In: *Weight, Motion and Force: Conceptual Structural Changes in Ancient Knowledge as a Result of its Transmission*. Unter Mitarb. von Jürgen Renn u. a. Preprints des Max-Planck-Instititutes für Wissenschaftsgeschichte. 320, 2006. URL: www.mpiwg-berlin.mpg.de/Preprints/P320.PDF.

[204] Jürgen Renn u. a. “Netzwerke als Wissensspeicher :” In: *Die Zukunft der Wissensspeicher : Forschen, Sammeln und Vermitteln im 21. Jahrhundert*. Hrsg. von J. Mittelstraß und U. Rüdiger. Konstanzer Wissenschaftsforum. München: UVK Verlagsgesellschaft Konstanz, 2016, S. 35–79. ISBN: 978-3-86764-716-8. URL: http://hdl.handle.net/11858/00-001M-0000-002B-61EE-C.

[205] *Researchspace: ResearchSpace Platform*. ResearchSpace, 29. Mai 2018. URL: https://github.com/researchspace/researchspace (besucht am 10. 06. 2018).

[206] *RFC 2518 - HTTP Extensions for Distributed Authoring WEBDAV*. URL: http://www.rfc-base.org/rfc-2518.html (besucht am 09. 06. 2018).

[207] *RFC 4511 - Lightweight Directory Access Protocol (LDAP): The Protocol*. URL: http://www.rfc-base.org/rfc-4511.html (besucht am 09. 06. 2018).

[208] Hans-Jörg Rheinberger. *Historische Epistemologie zur Einführung*. OCLC: 184000587. Hamburg: Junius, 2007. ISBN: 978-3-88506-636-1.

[209] Shawna Ross. “In Praise of Overstating the Case: A Review of Franco Moretti, Distant Reading (London: Verso, 2013)”. In: 8.1 (2014). URL: http://www.digitalhumanities.org/dhq/vol/8/1/000171/000171.html (besucht am 11. 05. 2016).

[210] M. Rosvall, D. Axelsson und C. T. Bergstrom. “The Map Equation”. In: *The European Physical Journal Special Topics* 178.1 (1. Nov. 2009), S. 13–23. ISSN: 1951-6355, 1951-6401. DOI: 10.1140/epjst/e2010-01179-1.

[211] Martin Rosvall und Carl T. Bergstrom. “Maps of Random Walks on Complex Networks Reveal Community Structure”. In: *Proceedings of the National Academy of Sciences* 105.4 (29. Jan. 2008), S. 1118–1123. ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.0706851105. pmid: 18216267.

[212] Antonio Sanfilippo u. a. “Automating Ontological Annotation with WordNet”. In: 2006.

[213] Maximilian Schich. “Figuring Out Art History”. In: (22. Okt. 2015). arXiv: 1512.03301 [physics, q-bio]. URL: http://arxiv.org/abs/1512.03301 (besucht am 15. 01. 2016).

[214] Maximilian Schich. *Rezeption und Tradierung als komplexes Netzwerk. Der CENSUS und visuelle Dokumente zu den Thermen in Rom*. 2007. URL: http://archiv.ub.uni-heidelberg.de/artdok/volltexte/2009/700 (besucht am 13. 12. 2015).

[215] M. J. Schiefsky. “New Technologies for the Study of Euclid’s Elements”. In: Euclid and His Heritage: Held at Oxford University. 2007. URL: http://archimedes.fas.harvard.edu/euclid/euclid_paper.pdf.

[216] Bernhard Schiemann u. a. *Erlangen CRM OWL*. URL: http://erlangen-crm.org/ (besucht am 25. 11. 2017).

[217] Urs Schoepflin. “The Archimedes Project: Realizing the Vision of an Open Digital Research Library for the Study of Long-Term Developments in the History of Mechanics”. In: *Proc. of the 5th National Russian Research Conference “Digital Libraries: Advanced Methods and Technologies, Digital Collections” RCDL*. 2003, S. 124–129.

[218] Matthias Schubert. *Datenbanken Theorie, Entwurf und Programmierung relationaler Datenbanken*. Hrsg. von SpringerLink (Online service). 2., überarbeitete Auflage. Wiesbaden: Teubner, 2007. ISBN: 978-3-8351-9108-2. URL: http://dx.doi.org/10.1007/978-3-8351-9108-2?nosfx=y.

[219] Feng Shi, Jacob G. Foster und James A. Evans. “Weaving the Fabric of Science: Dynamic Network Models of Science’s Unfolding Structure”. In: *Social Networks* 43 (Okt. 2015), S. 73–85. ISSN: 0378-8733. DOI: 10.1016/j.socnet.2015.02.006.

[220] *Siena Homepage*. URL: https://www.stats.ox.ac.uk/~snijders/siena/ (besucht am 10. 04. 2016).

[221] Tom A. B. Snijders. “The Multiple Flavours of Multilevel Issues for Networks”. In: *Multilevel Network Analysis for the Social Sciences*. Methodos Series. Springer, Cham, 2016, S. 15–46. ISBN: 978-3-319-24518-8 978-3-319-24520-1. DOI: 10.1007/978-3-319-24520-1_2. URL: https://link.springer.com/chapter/10.1007/978-3-319-24520-1_2 (besucht am 26. 08. 2017).

[222] Tom A. B. Snijders. “The Statistical Evaluation of Social Network Dynamics”. In: *Sociological Methodology* 31.1 (2001), S. 361–395. ISSN: 1467-9531. DOI: 10.1111/0081-1750.00099.

[223] *SPARQL 1.1 Property Paths*. URL: https://www.w3.org/TR/sparql11-property-paths/ (besucht am 30. 07. 2018).

[224] *SPARQL Endpoints for the Authority Files of the German National Library*. URL: http://wisski1.wiss-ki.eu/authorities/gnd/ (besucht am 10. 06. 2018).

[225] Steffen Staab und Rudi Studer, Hrsg. *Handbook on Ontologies*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. ISBN: 978-3-540-70999-2. URL: http://www.springerlink.com/content/978-3-540-70999-2/#section=675559&page=1 (besucht am 22. 10. 2010).

[226] Juliane Stiller und Dirk Wintergrün. “Digital Reconstruction in Historical Research and Its Implications for Virtual Research Environments”. In: 2016.

[227] Juliane Stiller u. a. “Anforderungen ermitteln, Lösungen evaluieren und Erfolge messen - Begleitforschung in DARIAH-DE.” In: *Bibliothek. Forschung und Praxis* 40.2 (2016), S. 250–258.

[228] Juliane Stiller u. a. *Nutzungsverhalten in den Digital Humanities :* 2015. URL: https://wiki.de.dariah.eu/download/attachments/14651583/Report1.2.1-final3.pdf.

[229] Susan E. F. Chipman und Christiane Fellbaum. “WordNet : An Electronic Lexical Resource”. In: (2017).

[230] TaDiRAH. *TaDiRAH - Taxonomy of Digital Research Activities in the Humanities*. 18. Juli 2014. URL: http://tadirah.dariah.eu/vocab/sobre.php (besucht am 15. 10. 2017).

[231] *TEI: P5 Guidelines*. URL: http://www.tei-c.org/Guidelines/P5/ (besucht am 18. 10. 2017).

[232] *TEI: Text Encoding Initiative*. URL: http://www.tei-c.org/index.xml (besucht am 18. 10. 2017).

[233] *TensorFlow*. URL: https://www.tensorflow.org/ (besucht am 19. 07. 2018).

[234] *Tesseract: Tesseract Open Source OCR Engine (Main Repository)*. tesseract-ocr, 20. Juli 2018. URL: https://github.com/tesseract-ocr/tesseract (besucht am 20. 07. 2018).

[235] *The Archimedes Project*. URL: http://archimedes2.mpiwg-berlin.mpg.de/archimedes_templates (besucht am 18. 10. 2017).

[236] *The Chymistry of Isaac Newton Project: Home*. URL: http://webapp1.dlib.indiana.edu/newton/ (besucht am 18. 10. 2017).

[237] *The Dataverse Project - Dataverse.Org*. URL: https://dataverse.org/ (besucht am 09. 06. 2018).

[238] *The Project | Islamic Scientific Manuscripts Initiative*. URL: https://ismi.mpiwg-berlin.mpg.de/ (besucht am 10. 06. 2018).

[239] *The Virtual Laboratory*. URL: http://vlp.mpiwg-berlin.mpg.de/index_html (besucht am 18. 10. 2017).

[240] *The Web Framework for Perfectionists with Deadlines | Django*. URL: https://www.djangoproject.com/ (besucht am 09. 06. 2018).

[241] Antonio Toral u. a. “Rejuvenating the Italian WordNet: Upgrading, Standardising, Extending”. In: *Proceedings of the 5th International Conference of the Global WordNet Association (GWC-2010)*. Mumbai, 2010.

[242] *Transititvity - Igraph R Manual Pages*. URL: http://igraph.org/r/doc/transitivity.html (besucht am 04. 09. 2017).

[243] Peter Turchin und S. A Nefedov. *Secular Cycles*. OCLC: 897048482. Princeton, N.J.: Princeton University Press, 2009. ISBN: 978-0-691-13696-7 978-1-4008-3068-8. URL: http://site.ebrary.com/id/10481991 (besucht am 12. 07. 2016).

[244] Can Türker. *Objektrelationale Datenbanken Ein Lehrbuch*. Bearb. von Gunter Saake. Bd. 1. Aufl.. Heidelberg: dpunkt-Verl., 2006. ISBN: 3-89864-190-2.

[245] John Unsworth. *Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?* 2000. URL: http://www3.isrl.illinois.edu/~unsworth/Kings.5-00/primitives.html (besucht am 20. 04. 2011).

[246] Thomas W. Valente. *Network Models of the Diffusion of Innovations*. Cresskill, NJ: Hampton Press, 1995.

[247] Matteo Valleriani. “The Tracts on the Sphere: Knowledge Restructured Over a Network”. In: *The Structures of Practical Knowledge*. Springer, Cham, 2017, S. 421–473. ISBN: 978-3-319-45670-6 978-3-319-45671-3. DOI: 10.1007/978-3-319-45671-3_16. URL: https://link.springer.com/chapter/10.1007/978-3-319-45671-3_16 (besucht am 20. 07. 2018).

[248] *Version 2.4 | FRBRoo*. web_version_frbroo. URL: http://www.cidoc-crm.org/frbroo/ModelVersion/version-2.4 (besucht am 05. 08. 2017).

[249] *Versions | CRMinf*. URL: http://new.cidoc-crm.org/crminf/fm_releases (besucht am 30. 10. 2017).

[250] *VirtualGraphs - Blazegraph*. URL: https://wiki.blazegraph.com/wiki/index.php/VirtualGraphs (besucht am 26. 11. 2017).

[251] *Virtuoso Graph Groups*. URL: http://vos.openlinksw.com/owiki/wiki/VOS/VirtRDFGraphsSecurity (besucht am 26. 11. 2017).

[252] *Visone*. URL: https://visone.info/ (besucht am 09. 06. 2018).

[253] Ulrike von Luxburg. “A Tutorial on Spectral Clustering”. In: *Statistics and Computing* 17.4 (1. Dez. 2007), S. 395–416. ISSN: 1573-1375. DOI: 10.1007/s11222-007-9033-z.

[254] *W3C XML Schema Definition Language (XSD) 1.1 Part 1: Structures*. URL: https://www.w3.org/TR/xmlschema11-1/ (besucht am 30. 10. 2017).

[255] H. Wan u. a. “Logic Programming with Defaults and Argumentation Theories”. In: *Logic Programming* (2009), S. 432–448.

[256] Peng Wang u. a. “Exponential Random Graph Models for Multilevel Networks”. In: *Social Networks* 35.1 (1. Jan. 2013), S. 96–115. ISSN: 0378-8733. DOI: 10.1016/j.socnet.2013.01.004.

[257] Wei Wang, Eric J. Neuman und Daniel A. Newman. “Statistical power of the social network autocorrelation model”. In: *Social Networks* 38.1 (2014), S. 88–99. ISSN: 0378-8733. DOI: 10.1016/j.socnet.2014.03.004.

[258] Duncan J. Watts und Steven H. Strogatz. “Collective Dynamics of ‘Small-World’ Networks”. In: *Nature* 393.6684 (4. Juni 1998), S. 440–442. ISSN: 0028-0836. DOI: 10.1038/30918.

[259] Hartmut Wedekind. “Are the Terms “Version” and “Variant” Orthogonal to One Another?” In: *SIGMOD Rec.* 23.4 (1994), S. 3–7.

[260] *Welcome To UML Web Site!* URL: http://www.uml.org/ (besucht am 30. 10. 2017).

[261] *Welcome to Zope Project and Community — Zope Project and Community Documentation*. URL: http://www.zope.org/en/latest/ (besucht am 09. 06. 2018).

[262] Marcus K. Weldon. *The Future X Network : A Bell Labs Perspective*. 2016.

[263] *Why Use DDI? | Data Documentation Initiative*. URL: https://www.ddialliance.org/training/why-use-ddi (besucht am 15. 10. 2017).

[264] *Wikidata*. URL: https://www.wikidata.org/wiki/Wikidata:Main_Page (besucht am 10. 06. 2018).

[265] *Wikidata Query Service*. URL: https://query.wikidata.org/ (besucht am 15. 08. 2017).

[266] *Wikidata:SPARQL Query Service/Wikidata Query Help - Wikidata*. URL: https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/Wikidata_Query_Help (besucht am 10. 06. 2018).

[267] Dirk Wintergrün. *Duomo Ontology*. URL: https://gitlab.gwdg.de/dirk.wintergruen/duomoOntology (besucht am 20. 08. 2018).

[268] Dirk Wintergrün. *duomoTools*. URL: https://gitlab.gwdg.de/MPIWG/Department-I/duomoTools (besucht am 10. 08. 2018).

[269] Dirk Wintergrün. *GraphExtensions*. URL: https://gitlab.gwdg.de/dirk.wintergruen/GraphExtensions (besucht am 10. 08. 2018).

[270] Dirk Wintergrün. *Schema Der Duomo DB*. URL: https://gitlab.gwdg.de/MPIWG/Department-I/duomoTools/raw/master/schemata/LEXXDUMP.rng (besucht am 10. 08. 2018).

[271] Dirk Wintergrün. *SPARQLGraph*. URL: https://gitlab.gwdg.de/dirk.wintergruen/SPARQLGraph (besucht am 10. 08. 2018).

[272] Dirk Wintergrün. *Src/Lex2skos.Py · Master · Dirk.Wintergruen / duomoPythonTools*. URL: https://gitlab.gwdg.de/dirk.wintergruen/duomoPythonTools/blob/master/src/lex2skos.py (besucht am 01. 05. 2019).

[273] Elisabeth Wolf und Iryna Gurevych. “Aligning Sense Inventories in Wikipedia and WordNet”. In: *Proceedings of the First Workshop on Automated Knowledge Base Construction*. Mai 2010, S. 24–28.

[274] *WordNet 2.0*. URL: https://www.w3.org/2006/03/wn/wn20/instances/index.html (besucht am 09. 08. 2017).

[275] *WordNet-Wikipedia: UKP*. URL: https://www.ukp.tu-darmstadt.de/data/lexical-resources/sense-alignment-resources/wordnet-wikipedia/ (besucht am 12. 08. 2017).

[276] Aleš Žiberna. “Generalized Blockmodeling of Valued Networks”. In: *Social Networks* 29.1 (Jan. 2007), S. 105–126. ISSN: 03788733. DOI: 10.1016/j.socnet.2006.04.002. arXiv: 1312.0646.

[277] Aleš Žiberna und Emmanuel Lazega. “Role Sets and Division of Work at Two Levels of Collective Agency: The Case of Blockmodeling a Multilevel (Inter-Individual and Inter-Organizational) Network”. In: *Multilevel Network Analysis for the Social Sciences*. Methodos Series. Springer, Cham, 2016, S. 173–209. ISBN: 978-3-319-24518-8 978-3-319-24520-1. DOI: 10.1007/978-3-319-24520-1_8. URL: https://link.springer.com/chapter/10.1007/978-3-319-24520-1_8 (besucht am 20. 08. 2017).