Kapitel 4 – Forschungsdesign

Ich verfolge den Ansatz einer verteilten Analyse von Zwischenräumen, in dem ich die Emergenz und die Bedeutungszuschreibungen der Datenwissenschaften in verschiedenen sozialen Feldern untersuche. In seinem Aufbau folgt das Kapitel der Grundstruktur der Arbeit: Zunächst ordne ich die Arbeit in das allgemeine Forschungsprogramm der Feldanalyse ein und erläutere das Forschungsdesign, um die Fragestellungen adäquat untersuchen zu können (Kap. 4.1). Danach beschreibe ich die Datengrundlage und das methodische Vorgehen für die drei empirischen Kapitel: erstens die Erhebung und Auswertung der Stellenanzeigen im Arbeitsmarkt (Kap. 4.2), zweitens die Strategiedokumente im Feld der Politik sowie das inhaltsanalytische Vorgehen (Kap. 4.3) und drittens Curricula sowie Interviews mit Lehrenden der Datenwissenschaften im akademischen Feld, die ich ebenfalls inhaltsanalytisch auswerte (Kap. 4.4). Abschliessend fasse ich das empirisch-analytische Potenzial des gewählten Forschungsdesigns hinsichtlich der verschiedenen Untersuchungsebenen zusammen (Kap. 4.5).

4.1 Die verteilte Analyse von Zwischenräumen

Die Arbeit verortet sich im »Forschungsprogramm« der Feldanalyse (Bernhard & Schmidt-Wellenburg 2012), das über den heterogenen Begriff des Feldes ein breites Spektrum empirisch-methodischer Vorgehensweisen abdeckt. Bernhard und Schmidt-Wellenburg (2012: 28) schlagen in Anlehnung an Lakatos (1978) vor, die Methodologie der Feldanalyse in einen programmatischen »Kern« von grundlegenden, stillschweigend vorausgesetzten Annahmen und einen flexibleren »Schutzgürtel«, der regelmässig empirischer Kritik und Auseinandersetzung unterzogen wird, zu differenzieren: Zum Kern des feldanalytischen Forschungsprogramms zählen die Autoren etwa die Einsicht, dass die soziale Realität als Ansammlung einander überlagernder und konkurrierender sozialer Felder zu verstehen und zu untersuchen ist. Dagegen sei im Schutzgürtel die Frage offen, »wie diese Felder zugeschnitten sind, welche Auseinandersetzungen sie im innersten bewegen und wie sie sich über die Zeit entwickeln« (Bernhard & Schmidt-Wellenburg 2012: 30f.).

Ich folge dieser forschungsprogrammatischen Unterscheidung, da das vorgeschlagene analytische Modell damit kompatibel ist: Räume zwischen Feldern als analytisches Konstrukt stützen sich elementar auf das Feldkonzept, erweitern es jedoch in einem zentralen Punkt, nämlich in der Frage der Lokalisierung neuer Wissensgebiete,

die nicht relativ autonome Felder, sondern unorganisierte, durchlässige und hybride Sphären sind. Daraus resultiert allerdings eine empirisch-methodische Herausforderung: Da die Feldtheorie Bourdieus, in geringerem Masse auch jene von Fligstein und McAdam, keinen ›Platz‹ für Zwischenräume als soziale Phänomene hat, finden sich in der feldanalytischen Methodendiskussion entsprechend keine Hinweise für deren Analyse (Bernhard & Schmidt-Wellenburg 2012; Blasius et al. 2019; Fligstein & McAdam 2012; Schmidt-Wellenburg & Bernhard 2020b).[1] Wie also können Zwischenräume als analytische Erweiterungen des Feldansatzes empirisch erforscht werden?

Eyal arbeitet in seinen Analysen zur Genese von Räumen zwischen Feldern bzw. Netzwerken der Expertise primär historisch, d. h., er rekonstruiert die Entwicklung solcher Wissenskonfigurationen durch eine umfassende Genealogie verschiedener Phasen, darin erscheinender, konkurrierender Expertisen und veränderter Akteurskonstellationen (Eyal 2002, 2013a, 2013b; Eyal & Pok 2015). Eine historische Rekonstruktion *post hoc* ist allerdings für die vorliegende Arbeit aus zwei Gründen keine Option. Erstens befindet sich das Wissensgebiet der Datenwissenschaften nach wie vor im Prozess seiner Etablierung – es ist noch unklar, wohin die Entwicklung führen wird: Etabliert es sich als ein Feld mit eigenen Spielregeln und Logiken, nach denen feldspezifisches Kapital zuteilwird? Existiert es auf Dauer als ein wenig regulierter, durchlässiger Raum zwischen Feldern? Oder verfällt es gar als eigener Bereich und existiert in anderen sozialräumlichen Phänomenen weiter? Zweitens ist das Phänomen in seiner Breite schlicht zu umfassend, als dass diese Arbeit eine vollständige Analyse der involvierten Felder wie Wissenschaft, Wirtschaft oder Politik leisten kann – vielmehr müssen dies spezifische Feldanalysen tun (Brandt 2016; Grommé et al. 2018).

Ich lege in dieser Arbeit deshalb einen anderen Fokus: Wie bereits ausgeführt, interessiert mich vor allem das feldübergreifende Moment des Phänomens, also der Umstand, dass Akteur*innen in verschiedenen Feldern an der Konstruktion der Datenwissenschaften beteiligt sind. Ich verfolge deshalb den Ansatz einer verteilten Analyse von Zwischenräumen[2]: Dabei untersuche ich ausgewählte Ausschnitte von Feldern im Hinblick auf das interessierende transversale Phänomen, die Emergenz der Datenwissenschaften. Dies ist möglich, da sich die Wirkungen der feldspezifischen Praktiken – d. h. die empirisch beobachtbaren »Feldeffekte« (Bourdieu & Wacquant 1996: 131) – nicht nur innerhalb der jeweiligen Felder der Wissenschaft, Ökonomie oder Politik, sondern ebenso im dazwischenliegenden Raum verorten lassen.

Mit der empirischen Vorgehensweise einer verteilten Analyse von Räumen zwischen Feldern schliesse ich an einen Kerngedanken der *multi-sited ethnography* bei George Marcus (1995; Nadai & Maeder 2005) an.[3] Marcus plädiert dafür, von einer

1 Bourdieu hat ein klares Vorgehen in der Analyse sozialer Felder formuliert (Bourdieu & Wacquant 1996: 136): Erstens gilt es, die Position eines Feldes im Verhältnis zum Feld der Macht zu untersuchen; zweitens um die Ermittlung der objektiven Struktur der Relationen zwischen den Positionen der in diesem Feld miteinander konkurrierenden Akteur*innen; drittens um die Analyse der Habitus, in denen die Dispositionen der Akteur*innen verinnerlicht sind.

2 Die Feldanalyse im Anschluss an Bourdieu schlägt eine ähnliche Vorgehensweise für transnationale Felder (Schmidt-Wellenburg & Bernhard 2020b: 2) bzw. transversale Felder (Witte & Schmitz 2019) vor.

3 Scheel et al. (2020) sprechen in ihrem ähnlich gelagerten methodischen Vorgehen von einer »transversalen Ethnographie«, was allerdings etwas irreführend ist in Bezug auf den Untersuchungsgegenstand, nämlich transversale Wissensfelder.

Fokussierung auf stationäre lokale Subkulturen und deren Einbettung in ein übergeordnetes Makrosystem wegzukommen und überkommene Dichotomien wie global vs. lokal, System vs. Lebenswelt etc. zu verwerfen. Vielmehr gelte es, die verteilten Phänomene in ihrer wechselseitigen Durchdringung zu betrachten, um fragmentierte, komplexe soziale Beziehungen erschliessen zu können (Marcus 1995: 98).

Eine solche Konzeption bietet sich analog für die Analyse von Zwischenräumen an: Anstelle einer monolithischen Analyse einzelner Felder, die binäre Strukturlogiken (dominant vs. dominiert, legitimes vs. illegitimes Kapital, autonom vs. heteronom etc.) reproduziert, strebt die Arbeit danach, den jeweiligen Beitrag der involvierten sozialen Felder bei der Herausbildung eines neuen Wissensgebietes zu betrachten, das eben gerade nicht ausschliesslich wissenschaftlich, ökonomisch, kulturell etc. ist. Vielmehr ist der Gegenstand in Zwischenräumen verteilt und fragmentiert zugleich: Es manifestieren sich in ihm Effekte verschiedener Felder, da er zum Objekt von konflikthaften wie kooperativen Praktiken – feldanalytisch formuliert quasi zum ›Spielball‹ – dieser Felder wird. Dennoch hat kein Feld das Primat oder die Deutungshoheit inne, sondern die Bedeutung ergibt sich erst durch die Gesamtheit der multiplen Perspektiven und Praktiken, die im Zwischenraum koexistieren. Es geht deshalb auch nicht darum, einen Vergleich der einzelnen Felder im Hinblick auf das interessierende Phänomen anzustellen, sondern darum, die Analyseteile des Puzzles zusammenzufügen (Nadai & Maeder 2005: 8). Dies erfolgt letztlich in der Absicht, mehr Erkenntnisse über das Phänomen zu gewinnen als die Summe der Analysen seiner einzelnen Teile – ein Postulat, das auch die Feldanalyse teilt (Bernhard & Schmidt-Wellenburg 2012: 35).

Von den unterschiedlichen Strategien der Feldkonstruktion, die Marcus in seinem Aufsatz vorschlägt, schliesst sich die vorliegende Arbeit jener der »Follow the Metaphor« (Marcus 1995: 108) an: Dabei leitet die Produktion und Zirkulation von Zeichen, Symbolen und anderen Repräsentationen in verschiedenen Bereichen das Forschungsdesign und empirische Vorgehen an. Marcus bezieht sich dabei prominent auf die Arbeit von Emily Martin (1993), die in ihrer Studie den Metaphern bzw. Konstruktionen folgt, wie in verschiedenen gesellschaftlichen Sphären über das Immunsystem gesprochen wird, von den Massenmedien über die ›Strasse‹, AIDS-Therapien, Alternativmedizin bis hin zur Wissenschaft.[4]

Ich gehe in der Arbeit ähnlich vor, um die Genese neuer Wissensgebiete zu erklären: Dazu folge ich gewissermassen den Effekten von Begriffs- und Grenzarbeit, durch die in unterschiedlichen Feldern multiple Bedeutungen des Gegenstandes Datenwissenschaften entworfen werden. Somit bietet es sich an, den Untersuchungsgegenstand in den relevanten Feldern verteilt zu untersuchen. Dabei verfolge ich einen stärker akteursorientierten Ansatz: Repräsentationen eines (neuen) Gegenstandes zirkulieren nicht einfach so in sozialen Feldern. Sie werden primär von Organisationen, d. h. kollektiven Akteur*innen, mobilisiert, die damit spezifische Ziele und Interessen in ihren jeweiligen Feldern verfolgen. Sie versuchen eine bestimmte Deutung des Gegenstandes zu etablieren, befördern allerdings synchron dazu auch die Entstehung neuer Möglichkeitsräume, um dadurch ihre Handlungsoptionen zu erweitern.

Um die Herausbildung von Zwischenräumen empirisch-analytisch untersuchen zu können, schlage ich ein dreistufiges Vorgehen vor: Erstens gilt es die relevanten umgebenden Felder zu identifizieren. In Anknüpfung an die (unscharfen) Grenzen

4 Eine ähnliche Strategie schlägt Abbott (2005: 249) für das vielschichtige Phänomen Alkoholismus vor.

von Feldern stellt sich die Frage: Die Effekte welcher Felder sind im Zwischenraum zu beobachten? An die Auswahl der umgebenden Felder anknüpfend ergibt sich zweitens die Frage nach den Akteur*innen: Welche Akteur*innen dieser Felder sind wie an der Konstruktion solcher Räume beteiligt? Inwiefern tragen sie zur Herausbildung neuer Räume bei? Sind die Relationen zwischen den Akteur*innen eher konflikthaft oder eher kooperativ? In einem dritten Schritt stellen sich Fragen nach den fundierenden Praktiken, die zur Etablierung von Zwischenräumen beitragen: Welche Praxismodi, d. h. welche kollektiven Stellungnahmen sind zu beobachten? Auf welchen Ebenen, d. h. wo finden diese statt bzw. von wo werden sie geäussert?

Die Auswahl der relevanten Felder leitet sich primär aus dem Forschungsstand ab, in dem ich die Bedeutung der Wissenschaft, der Ökonomie sowie der Politik als relevante soziale Sphären in der Konstruktion der Datenwissenschaften herausgearbeitet habe. Im Feld der Wirtschaft werden einerseits hohe Erwartungen an Data Scientists als Schlüsselfiguren datengetriebener Produktionsweisen formuliert, die insbesondere in der Hochschulbildung stark rezipiert werden (Saner 2019), andererseits zeigen sich manifeste Unsicherheiten und divergierende Deutungen über die relevanten Kategorien und Inhalte. In der Wissenschaft beschäftigen sich verschiedene Disziplinen seit Jahrzehnten mit den konkreten Praktiken und Wissensinhalten der Datenwissenschaften. Die Transformation hin zu einer datengetriebenen Wissensproduktion reartikuliert dabei existierende Konflikte und Grenzziehungen. Gleichzeitig eröffnen sich Opportunitäten, solche Konfliktlinien durch grenzüberschreitende Kooperationen in Forschung und Lehre zu überwinden und heterogene Perspektiven auf die Datenwissenschaften zu etablieren. Schliesslich prägen Akteur*innen der Hochschul- und Forschungspolitik Szenarien einer vielversprechenden Zukunft, die durch Investitionen und Fördermassnahmen begleitet werden. Somit koexistieren in den drei zentralen Feldern multiple Deutungen und konturieren über ihre Effekte den Raum dazwischen.

Geographisch beschränke ich mich dabei auf die Herausbildung der Datenwissenschaften und die Konfiguration der relevanten umgebenden Felder in der Schweiz. Obwohl neue Wissensfelder oft global strukturiert sind, bleiben sie durch lokale, regionale und nationalstaatliche Spezifika gekennzeichnet (Biniok 2013; Merz & Sormani 2016). Dennoch versuche ich, die transnationalen Relationen und wechselseitigen Abhängigkeiten mit zu berücksichtigen, um die Räume zwischen Feldern nicht als nationalstaatlich verfasst zu reifizieren (Schmidt-Wellenburg & Bernhard 2020b).

Um die in der Einleitung genannten Forschungsfragen empirisch-methodisch umsetzen zu können, verknüpfe ich das für die Soziologie relativ ›neue‹ Verfahren des Topic Modeling (DiMaggio et al. 2013; Mützel 2015a; Papilloud & Hinneburg 2018) mit ›traditionellen‹ sozialwissenschaftlichen Methoden der qualitativen Inhalts-, Curricula- und Interviewanalyse. Zunächst ermöglicht es das Topic Modeling von Stellenanzeigen, transversale Repräsentationen der Datenwissenschaften auf der gesellschaftlichen Makroebene zu analysieren. Obwohl hier das Material feldspezifisch nicht begrenzt ist, zeigt sich bei der Datenvorbereitung, dass Akteur*innen im ökonomischen Feld zentral sind. Das gewählte methodische Vorgehen ist jedoch in der Lage, sowohl feldspezifische als auch feldübergreifende Darstellungen zu identifizieren und letztere in allgemeinen Topics zu modellieren.

Im Fall des Politikfeldes bilden Strategiedokumente unterschiedlicher Akteur*innen der Hochschul- und Forschungspolitik das Datenmaterial. Durch eine qualitative Inhaltsanalyse können die spezifischen Zukunftsvisionen, die im Rahmen des Diskur-

ses über die Digitalisierung auf die Datenwissenschaften entworfen werden, rekonstruiert werden. Effekte dieser kollektiven Praktiken sind nicht nur in der Politik selbst, sondern auch in anderen Feldern, insbesondere der Wissenschaft und der Ökonomie zu beobachten.

Schliesslich bilden Curricula aller verfügbaren Studienangebote in Datenwissenschaften an Schweizer Universitäten und Hochschulen sowie Interviews mit Lehrenden das empirische Untersuchungsmaterial im akademischen Feld, die ich ebenfalls mittels qualitativer Inhaltsanalysen analysiere. Dies erlaubt es zum einen, die fundierende Rolle von Praktiken der Begriffsarbeit in der Herausbildung des Zwischenraumes als eines Möglichkeitsraums zu bestimmen. Zum anderen bietet die Kombination von kollektiven (Curricula) und subjektiven Stellungnahmen (Interviews) die Möglichkeit, die Widersprüche und Dynamiken synchroner Praktiken der Grenzziehung und Grenzüberschreitung zu adressieren.

4.2 Stellenanzeigen im Arbeitsmarkt

Der Forschungsstand hat deutlich gemacht, dass in der Suche und Diskussion über die relevanten Kompetenzen ein zentrales Moment in der Konstruktion der Datenwissenschaften liegt. Durch die Erhebung von Stellenanzeigen für Data Scientists können divergierende Repräsentationen der Datenwissenschaften in Stellenanzeigen feldübergreifend analysiert werden. Um ein Sample von Stellenausschreibungen für Data Scientists und verwandten Bezeichnungen zusammenstellen, aufbereiten und auswerten zu können, war ein mehrstufiges Verfahren notwendig, das ich hier ausführlich darlegen werde. Zunächst lege ich im ersten Abschnitt einige allgemeine Charakteristika von Stelleninseraten als Daten zugrunde (Kap. 4.2.1), bevor ich im zweiten Teil auf die Prozesse der Datenerhebung (Kap. 4.2.2) sowie der Kodierung und Datenbereinigung (Kap. 4.2.3) eingehe. Im dritten Teil erläutere ich schliesslich die Vorbereitung und Validierung der Texte (Kap. 4.2.4) sowie das methodisch-analytische Vorgehen mittels Topic-Modeling-Verfahren (Kap. 4.2.5).

4.2.1 Stellenanzeigen als Forschungsgegenstand

Stelleninserate sind in der Soziologie ein vergleichsweise wenig beachtetes empirisches Material (Geser 1983; Kriesi et al. 2010; Salvisberg 2010; Schreiber 1995). So enthält ein Standardwerk zur (deutschsprachigen) Arbeitsmarktsoziologie keinerlei Bezüge zu Stelleninseraten als Daten (Abraham & Hinz 2018). Dies ist umso erstaunlicher, da es sich um eine interessante Datenquelle handelt: Stelleninserate formulieren die unterschiedlichen Repräsentationen des Untersuchungsgegenstandes, etwa indem sie divergierende Tätigkeitsfelder, Kompetenzzuschreibungen oder Qualifikationsanforderungen nennen und einfordern. In solchen Konstruktionsleistungen verdichten sich berufsspezifische Rollenerwartungen (Geser 1983): Sie artikulieren feld- und organisationsspezifische Perspektiven darauf, worin bestimmte Praktiken und Expertisen bestehen und wo die Berührungspunkte sowie Differenzen zu verwandten Gebieten liegen.

Bei Stellenanzeigen handelt es sich um prozessgenerierte Daten, da sie im ordentlichen Handlungsfluss von Akteur*innen entstehen und entsprechend auf einem

»nicht-reaktiven Erhebungsprozess« basieren (Weischer 2015: 73). Sie folgen kulturellen und sprachlichen Konventionen, beispielsweise was die erwarteten Qualifikationsanforderungen oder Aussagen zur finanziellen Entlohnung betrifft (Geser 1983: 478), aber auch rechtlichen Regulierungen (Aratnam 2012). In der Regel weisen Stelleninserate einen stark standardisierten Aufbau auf: Sie enthalten Informationen über die inserierende Organisation (*wer sucht*), die Stelle und die zu erledigenden Aufgaben (*für was wird gesucht*), Anforderungen an Ausbildung, Erfahrung und Sachkenntnisse, zugeschriebene Eigenschaften und Fähigkeiten (»Kompetenzen«) oder Kriterien wie Alter und Geschlecht (*wer wird gesucht*) (Salvisberg 2008: 2567).[5] Abschliessend wird neben Bewerbungsfristen und Kontaktmöglichkeiten bisweilen auf weitere Faktoren wie das Arbeitsumfeld und Chancengleichheits- bzw. Diversity-Statements (vor allem in englischsprachigen Stelleninseraten) verwiesen.

Die Bestandteile von Stelleninseraten sind oft mit unterschiedlich strukturierten Textformaten verknüpft: Während die organisationalen Selbstbeschreibungen sowie die Tätigkeitsprofile in der Regel prosaisch in integralen Sätzen verfasst sind (vgl. die Abschnitte [1]–[5] im Beispielinserat im Anhang), werden die Anforderungen und Qualifikationen oft in Listen unvollständiger Satzkonstruktionen oder nur stichwortartig formuliert (vgl. den Abschnitt [6] »Ihr Profil« im Beispielinserat im Anhang). Auf der Ebene des Vokabulars äussert sich dies in spezifischen Kombinationen von organisationsbezogenen, feld- sowie stelleninseratetypischen Begriffen.

Als Spezifika von Stelleninseraten in technologiebezogenen Wissensgebieten wie Informatik oder Datenwissenschaften gelten in der fach- und berufspädagogischen Diskussion divergierende Annahmen über die Qualifikationsanforderungen von Bewerber*innen (Bott et al. 2000; Schumann et al. 2016; Wowczko 2015). Was für die Akteur*innen im Feld ein Problem darstellen mag, weil immer unklar bleibt, über welche Kompetenzen nun eine Datenwissenschaftlerin verfügen muss, ist für die empirische Analyse hingegen bereits eine erste konstitutive Erkenntnis: Die Debatten über die notwendigen Skillsets oder Methoden verweisen auf die Suche nach den ›richtigen Kompetenzen‹. Sie sind gewissermassen fundierend für die Herausbildung des Phänomens an sich. Weiter manifestiert sich in Stellenanzeigen die eminente Bedeutung unterschiedlicher Werkzeuge wie der verwendeten Betriebssysteme, Software, Programmiersprachen und weiterer Instrumente: So ist im empirischen Material zu beobachten, dass Methoden, Tools oder Qualifikationsanforderungen oft als Synonyme verwendet werden (»Mehrjährige Erfahrung mit statistischer Software, R, Python, SAS oder SPSS«, vgl. Beispielinserat 1). Solche nicht-hierarchisierenden Aufzählungen deuten eine gewisse Beliebigkeit im Sinne »unverbindliche[r] Wunschlisten« an (Geser 1983: 479).

Die empirische Aussagekraft der Datenquelle Stellenanzeigen wird insofern eingeschränkt, dass sie nur die formale, nachfrageorientierte Seite des Arbeitsmarktes abbilden, während informelle Wege (wie Praktika, Abschlussarbeiten, persönliche

5 Salvisberg (2010: 116f.) unterscheidet folgende acht Analyseeinheiten: Selbstbeschreibung des inserierenden Unternehmens; Angaben zum Grund der Stellenausschreibung; Informationen zur administrativen Abwicklung der Bewerbung; ggf. Angaben zur Stellenvermittlerin; Hinweise auf materielle Gegenleistungen des Arbeitgebers; Stellen- und Aufgabenbeschreibung; Anforderungen an Ausbildung, Erfahrung und Sachkenntnisse; notwendige Charaktereigenschaften, Sozialkompetenzen oder Zuschreibungskriterien wie Alter oder Geschlecht.

Netzwerke etc.) oder Spontanbewerbungen auf der Angebotsseite nicht berücksichtigt werden.[6] Auch organisationsinterne Arbeitsmärkte (Köhler & Krause 2010; Struck 2018) werden durch das gewählte Verfahren nicht erfasst. Nichtsdestotrotz bilden Stelleninserate aus den genannten Gründen einen interessanten Forschungsgegenstand, weil sie es ermöglichen, die multiplen Bedeutungen eines sozialen Phänomens über die unterschiedlichen sozialen Zuschreibungen, Qualifikationsanforderungen oder Tätigkeitsfelder zu rekonstruieren. Dies stellt insbesondere im Falle neuer Wissensgebiete bzw. solcher in Transformation eine Möglichkeit dar, prozessinduzierte Dynamiken jenseits subjektiver Stellungnahmen (wie Kommentare, Blogs, Tweets etc.) als auch reaktiver Erhebungsmethoden (wie Umfragen) zu untersuchen.

4.2.2 Datenerhebung

Die Stellenausschreibungen für Data Scientists und verwandte Bezeichnungen (wie Data Analyst, Data Engineer etc.) bilden die Datengrundlage für die Analyse im Kapitel 5. Um das Korpus zu bilden, habe ich auf der Schweizer Stellenplattform jobs.ch veröffentlichte Stellenbeschreibungen, die die Suchbegriffe »Data AND Scientist«, »Data AND Science« oder »Big AND Data« enthalten, mittels Webscraping-Verfahren (Munzert et al. 2015) in der Statistiksoftware R erfasst und als Textdateien gespeichert. Zwischen Januar 2017 und Juni 2019 konnte ich ein Sample von N = 4341 Stellenanzeigen zusammenstellen.[7] Die Suchbegriffe wurden bewusst kombiniert, um einerseits die Breite des Phänomens zu adressieren und andererseits keine Engführung auf bestimmte mit den Begriffen verknüpfte Orientierungen (beispielsweise »Big Data« als informatisches gegenüber »Data Science« als statistisches Konzept) zu riskieren.

Während des Erhebungsprozesses fiel auf, dass diverse inserierte Stellen über Monate, manchmal sogar über mehrere Jahre hinweg verfügbar blieben. Oft wurden die Inserate mehrmals leicht adaptiert und unter verändertem Berufstitel oder mit einem veränderten Anforderungsprofil erneut publiziert.[8] Dies verweist auf andauernde Schwierigkeiten in der Rekrutierung von Datenwissenschaftler*innen (Burtch 2016; H. D. Harris et al. 2013; Markow et al. 2017) sowie deren Konstruktion als stark nachgefragte Berufsgruppe.

Das Sample habe ich fortlaufend sowohl manuell als auch mittels automatisierter Texterkennung auf Duplikate überprüft, um Mehrfachzählungen derselben Inserate zu vermeiden. Dabei habe ich folgende Regel festgelegt: Verwenden zwei oder meh-

6 Aufgrund der sehr hohen Nachfrage und der Vorbereitung der Studierenden auf informelle Kontaktkanäle kommt diesen für das untersuchte Arbeitsmarktsegment eine gewisse Bedeutung zu (vgl. Gerlach 2014: 145 für das ähnlich gelagerte Beispiel von Automobilingenieur*innen).

7 Das Sample stellt keine Vollerhebung dar; es handelt sich um einen selektiven Ausschnitt aus dem Total aller Stellenanzeigen. Nimmt man den »Schweizer Jobradar« der Firma x28 AG, der in Anspruch nimmt, ein »vollständiges und repräsentatives Bild der Nachfrage nach Arbeitskräften« in der Schweiz zu liefern, als Vergleichsmass, lässt sich näherungsweise festhalten, dass das Sample durchschnittlich rund 2.4 ‰ der pro Quartal erfassten Stellenanzeigen entspricht (eigene Berechnungen). Vgl. online: https://www.x28.ch/jobradar/ (Zugriff: 03.02.2022).

8 Zudem ergab sich die Schwierigkeit, dass der HTML-Code der Webseite mehrmals geändert wurde. Dies führte dazu, dass unterschiedliche Stellenanzeigen in die Suchanfrage inkludiert wurden. Eine Änderung des HTML-Codes ist beispielsweise für Oktober 2017 dokumentiert, was sich in einem deutlich erkennbaren Anstieg der Anzahl erfasster Anzeigen niederschlägt (vgl. Abbildung 3).

rere Anzeigen einen identischen Stellentitel und Ausschreibungstext, aber ein unterschiedliches Datum, so werden diese aus dem Sample entfernt. Enthalten zwei Anzeigen einen identischen Titel, aber unterschiedliche Stellentexte (beispielsweise indem das Anforderungsprofil adaptiert wurde), oder umgekehrt variierende Stellenbezeichnungen, aber identische Stellentexte, so sind beide im Sample enthalten.[9]

4.2.3 Kodierung und Datenbereinigung

Eine zentrale Herausforderung von Daten, die durch soziotechnische Systeme (wie Social-Media-Plattformen und andere internetbasierte Dienste) prozessual generiert werden, ist, dass sie inhärent »messy« sind.[10] Für den ersten Erhebungs- und Auswertungsschritt der Arbeit ergab sich insofern das Problem, dass parallel zur Datenerhebung mehrere Teilschritte der Kodierung und Datenbereinigung notwendig wurden, um die Daten zu vereinheitlichen und so für die Analyse verfügbar zu machen.[11] Aufgrund der hohen Anzahl Stellenanzeigen im Sample sollte dies automatisiert erfolgen. So liessen sich aus den erfassten Textdateien automatisiert diverse Informationen extrahieren, die für die deskriptive Analyse relevant sind: Stellenbezeichnung, Datum, inserierende Organisation (Arbeitgeber), lokale Standorte, die URL sowie der Text der Anzeige. Die Attribute wurden als einzelne Variablen kodiert.[12] Deren Häufigkeiten und Verteilungen bilden die Grundlage für die deskriptive Analyse des Samples.

Um Informationen über die Akteur*innen der Datenwissenschaften in den einzelnen Feldern zu erhalten, mussten die Namen der Organisationen, die Stellenanzeigen publizieren, vereinheitlicht werden (Gross- und Kleinschreibung, mit oder ohne rechtsformindizierende Angaben wie AG, GmbH etc.). Anschliessend wurden diese nach ökonomischen bzw. anderen sozialen Feldern kodiert. Als Grundlage dazu dienten die Selbstbeschreibungen der Firmen auf ihren Webseiten. Die Kodierung der Unternehmen und Organisationen in ökonomische und andere Felder entspricht damit nicht wirtschaftsstatistischen Klassifikationen des Bundesamts für Statistik (BFS), sondern hat sich während des Kodierprozesses induktiv aus dem Material gebildet.[13] Die inserierenden Organisationen im Datensatz entstammen achtzehn unterschiedlichen Feldern.

---

9 Ein Beispiel: Die Firma Helbling AG publizierte ein Inserat für eine/n »Data Scientist (m/w) – Schwerpunkt Signalverarbeitung/Algorithmen für optische Messsysteme«, datiert auf den 11. Februar 2019. Dieses wurde rund einen Monat später, am 15. März 2019, unter der Bezeichnung »Entwicklungsingenieur (m/w) – Schwerpunkt Signalverarbeitung/Optik« über dieselbe URL erneut erfasst. Beide Inserate bleiben im Sample enthalten.

10 Der Umgang damit ist ein zentraler Aspekt datenwissenschaftlicher Praxis (Mützel et al. 2018: 114ff.).

11 Die bereinigten und kodierten Daten sind auf dem Datenrepositorium FORSbase hinterlegt (Saner & Mützel 2021a).

12 Zusätzlich wurde jeder Anzeige eine eigene Identifikationsnummer zugewiesen.

13 Bei der ersten Kodierung hatten Unternehmen im Feld der Personaldienstleistungen den höchsten Anteil. Eine detaillierte Betrachtung der Inserate zeigte, dass diese nur in der Minderheit der Fälle (<5 %) für sich selbst rekrutieren, sondern dies überwiegend im Auftrag anderer Firmen tun (indiziert durch Begriffe wie *client*, *customer*, *Kunde/Kundin oder Auftraggeber*in* im Inseratetext). Bei diesen handelte es sich mehrheitlich um die bedeutenden Felder der Pharmaindustrie sowie der Banken/FinTech (erschienen in je rund 34 % der Inserate). Auch Felder wie IT/Software (19 %), Industrie (13 %) oder Versicherungen (7 %) wurden oft genannt. In der Folge wurden alle Inserate von Personalrek-

Die Stelleninserate sind in vier verschiedenen Sprachen (Englisch, Deutsch, Französisch und Italienisch) publiziert. Computerlinguistische Verfahren der automatisierten Spracherkennung ermöglichen es hier, den Anzeigentext eindeutig einer der vier Sprachen zuzuordnen, was Analysen über die sprachräumliche Verteilung der Datenwissenschaften in der Schweiz erlaubt.[14] Zusätzlich stellt die Sprache einer Stellenanzeige einen Indikator für die Internationalität einer Organisation bzw. eines Feldes dar.

Auch die Standorte der inserierenden Unternehmen und Organisationen wurden vereinheitlicht (Entfernung von Doppelnamen wie Zürich-Oerlikon oder englischen Städtenamen [Geneva] in eindeutige Standortangaben), um diese eindeutig geographisch lokalisieren zu können. Anschliessend erfolgte die Einteilung der Standorte der Unternehmen und Organisationen in die sieben geographischen Grossregionen der Schweiz (Genferseeregion, Espace Mittelland, Nordwestschweiz, Zürich, Ostschweiz, Zentralschweiz sowie das Tessin) gemäss BFS.[15] Waren zwei oder mehr Standorte genannt, die sich über mehrere der genannten Grossregionen hinweg verteilen, wurde die Region als »Mehrfachnennung« kodiert. Rund ein Prozent der Inserate liess sich keinem eindeutigen Standort zuweisen oder dieser liegt ausserhalb der Schweiz.

4.2.4 Vorbereitung und Validierung der Daten für Text Mining

Die beschriebenen Informationen über die feldspezifische, geographische und sprachliche Verortung der Stelleninserate geben Hinweise über die Struktur und Entwicklung der Datenwissenschaften im schweizerischen Arbeitsmarkt. Das Haupterkenntnisinteresse der Untersuchung gilt allerdings den Ausschreibungstexten der Stelleninserate: In den Texten werden sehr unterschiedliche Tätigkeitsfelder, Kompetenzzuschreibungen oder Qualifikationsanforderungen entworfen, d. h. multiple Repräsentationen der Datenwissenschaften formuliert. Aufgrund ihrer »messyness« infolge des Webscraping habe ich die Texte zunächst manuell und anschliessend automatisiert überprüft und bereinigt, um eine möglichst hohe Datenqualität zu garantieren und ihre Interpretierbarkeit in textanalytischen Verfahren zu gewährleisten (Maier et al. 2018: 100f.).[16] Dies beinhaltete mehrere Schritte der Datenvorbereitung mittels des Quanteda Package in R (Benoit et al. 2018).[17]

---

rutierungsfirmen, die sich über den Text eindeutig einem bestimmten ökonomischen Feld zuordnen liessen, manuell rekodiert.

14 Das textcat Package in R (Hornik et al. 2013) verwendet dazu eine n-gram-basierte Textkategorisierung (Jurafsky & Martin 2018): Texte werden in einzelne Bestandteile zerlegt und sodann mittels Häufigkeitsanalysen der jeweiligen Sprache zugeordnet. Die Ergebnisse dieses Verfahrens wurden regelmässig auf ihre Validität geprüft.

15 Vgl. Bundesamt für Statistik (2020): Analyseregionen: https://www.bfs.admin.ch/bfs/de/home/statistiken/querschnittsthemen/raeumliche-analysen/raeumliche-gliederungen/analyseregionen.html (Zugriff: 03.02.2022).

16 Die Länge der in der Analyse verwendeten Dokumente weist einen Durchschnitt von 530 und einen Median von 488 Wörtern auf, wobei die englischsprachigen Dokumente (Durchschnitt = 581 Wörter, Median = 547 Wörter) etwas länger sind als die deutschsprachigen (Durchschnitt = 414 Wörter, Median = 381 Wörter).

17 Zu den Vorbereitungsschritten gehörten die Transformation aller Token in Kleinbuchstaben, das Entfernen von HTML-Tags, Satzzeichen sowie sprachenspezifischer Stoppwörter. Zusätzlich entfern-

Nach diesen Vorbereitungsschritten erstellte ich je eine Dokument-Feature-Matrix (DFM) für die englisch- sowie die deutschsprachigen Stelleninserate, wobei jedes Dokument einen eigenen Vektor darstellt. Die Texte werden in ihre einzelnen Bestandteile (Tokenisierung) zerlegt, d. h., die tatsächliche Reihenfolge der Wörter und damit auch deren syntaktische Struktur werden ignoriert. Dadurch wird auch der Umstand neutralisiert, dass es sich bei Stellenanzeigen um ein Textgenre handelt, das integrale Satzbestandteile mit Wortauflistungen kombiniert. Das Entfernen seltener Begriffe, die in weniger als fünf unterschiedlichen Dokumenten im Korpus erscheinen, reduziert die Dimensionalität der Matrizen um rund drei Viertel (-75.6 % im englischsprachigen bzw. -78.8 % im deutschsprachigen Korpus). Die resultierende englischsprachige DFM enthält 2941 Dokumente mit 5771 Features, die deutschsprachige DFM 1282 Dokumente mit 4671 Features.

Solche Schritte der Textvorbereitung machen den Datensatz schlanker und somit für die Analyse leichter handhabbar, reduzieren jedoch gleichzeitig auch die Menge an Informationen, die im Datensatz enthalten sind. Entsprechend muss jeder einzelne Schritt der Textvorbereitung sorgfältig überprüft und validiert werden (Denny & Spirling 2018; Maier et al. 2018). Insbesondere das Stemming (Stammformreduktion) von Wörtern, d. h. das regelbasierte Kürzen der Endungen von Wörtern auf ihren Stamm, kann zu sehr unterschiedlichen Ergebnissen beim Topic Modeling führen (Blei 2012; Denny & Spirling 2018; Maier et al. 2018). Denny und Spirling (2018) haben mit preText ein Tool entwickelt, mit dem die Effekte der einzelnen Textvorbereitungsschritte insbesondere für nicht-überwachte Formen maschinellen Lernens (*unsupervised learning*) evaluiert werden können. Je tiefer der resultierende Wert, desto weniger beeinflusst die Kombination der gewählten Textvorbereitungsschritte das Korpus. Trotz der Information, inwiefern ein bestimmtes Modell das Korpus beeinflusst, gibt der Wert in diesem Sinne keine finale Antwort für ein zu wählendes Modell, sondern schafft die Möglichkeit, die gewählte Variante im Vergleich zu anderen Modellen zu evaluieren. Die gewählten Varianten[18] erwiesen sich im Vergleich mit anderen evaluierten Modellen als robust und in der Analyse mittels Topic Modeling auch als tatsächlich interpretierbar.[19]

---

te ich Begriffe, die spezifisch für die erhobenen Stelleninserate sind (*jetzt, bewerben, Originalinserat, anzeigen*) oder solche, die aufgrund des Aufbaus und der Struktur der Webseite jobs.ch in die Datenerhebung einflossen (*bitte, beziehe, dich, bei, deiner, Bewerbung, auf, jobs.ch*), da sie für die inhaltliche Analyse keinen Informationswert besitzen, umgekehrt allerdings die Ergebnisse des Topic-Modeling-Verfahrens verzerren können (vgl. unten).

18 Die gewählte Variante mit dem Akronym P(unctuation)-N(umbers)-L(owercase)-S(tem)-W(Stopword Removal) weist für das englischsprachige Korpus mit preText = 0.132 den neuntniedrigsten Wert aller 128 möglichen Kombinationen auf (die Spannweite ist bei r = 0.225). Der Wert für das deutschsprachige Sample ist mit preText = 0.127 etwas niedriger. Dieser liegt zwar nur im Mittelfeld aller möglichen Kombinationen, wobei die Differenz zum ›besten‹ Wert lediglich 0.027, gegenüber dem ›schlechtesten‹ Wert jedoch 0.179 beträgt (die Verteilung der Werte ist stark linksschief).

19 Dies habe ich exemplarisch anhand der drei Topic-Modelle mit den niedrigsten preText-Werten für das englisch- und deutschsprachige Sample überprüft: Die Ergebnisse waren unter anderem aufgrund der Kombination von Bi- und Trigrammen kaum interpretierbar, da keine Stopwörter entfernt wurden.

4.2.5 Analytisches Vorgehen mittels Topic Modeling

Nun liegen die Textdateien als bereinigte DFM vor und können mittels textanalytischer Methoden (Evans & Aceves 2016; Ignatow & Radev 2016) untersucht werden. Die Analyse grosser Mengen an Textdaten erfährt gegenwärtig unter der Losung *Text as Data* (Grimmer & Stewart 2013) eine erneuerte Konjunktur in den Sozialwissenschaften. In einer kultursoziologischen Perspektive erweist sich insbesondere Topic Modeling als vielversprechendes Verfahren, um grosse Textdaten auf latente Muster und Strukturen hin zu untersuchen und mit der Analyse sozialer bzw. kultureller Felder zu verknüpfen (Bail 2014; DiMaggio et al. 2013; Mohr & Bogdanov 2013; Mützel 2015b; Papilloud & Hinneburg 2018). Topic Modeling umfasst computerlinguistische probabilistische Verfahren, um Textdokumente statistisch zu annotieren und somit ein Korpus dimensionenreduzierend zu strukturieren.

Bei Topic Modeling werden »aufgrund einer bestimmten Verteilungsannahme Themen (topics) aus Texten generiert. Topic Modeling Algorithmen analysieren Wörter aus einem Textkorpus, um darin enthaltene Themen zu identifizieren. Dies ermöglicht dann, Themen zueinander in Beziehung zu setzen und im zeitlichen Verlauf abzubilden« (Mützel 2015b: 411). Ein prominenter Algorithmus ist die »Latent Dirichlet Allocation« (LDA) (Blei et al. 2003): Das Topic-Modell verwendet Ko-Okkurrenzen von Wörtern, um latente Themen in den Dokumenten zu eruieren. Die Dokumente werden nicht als einem Thema alleine zugehörig betrachtet, sondern können verschiedenen Themen angehören. Gleichzeitig variieren die Verhältnisse der Themen über die Dokumente. So setzt sich ein bestimmtes Dokument beispielsweise zu 62 % aus Topic 1, zu 23 % aus Topic 3 und zu 15 % aus einer Mischung der restlichen Topics zusammen. Die Zahl der Topics muss zu Beginn manuell definiert werden. Ausgehend von dieser Definition berechnet LDA die Topics, die aus unterschiedlichen Themenvokabularen bestehen. Während also den Dokumenten Topics zugeordnet werden, zeichnen sich Topics durch eine charakteristische Begriffssammlung aus.

Der Nutzen einer Modellierung des Textkorpus mittels LDA liegt darin, dass die Multidimensionalität von Begriffen einer grossen Anzahl von Dokumenten automatisiert untersucht werden kann (DiMaggio et al. 2013): Begriffe wie »Daten«, »Methoden« oder »Wissenschaft« verfügen je nach Kontext über multiple Bedeutungen. Ein bestimmter Begriff kann in Kombination mit weiteren Begriffen eines Topics eine andere inhaltliche Bedeutung aufweisen wie derselbe Begriff in einem anderen Topic. Topic-Modelle sind in der Lage, die vielfältigen Konnotationen in unterschiedlichen Topics zu identifizieren und zu rekonstruieren. Dies erlaubt es, die multiplen Dimensionen und Bedeutungsebenen des Untersuchungsgegenstandes Datenwissenschaften zu analysieren.

Die latente Struktur der Topics eines Korpus berechne ich mittels Gibbs-Sampling (Griffiths & Steyvers 2004: 5229f.) und 2000 Iterationen: Dabei wird nach einer zufälligen Initialisierung die Topic-Zuordnung jedes einzelnen Wortes in allen Dokumenten in einem iterativen Verfahren wiederholt vorgenommen. Die Neuzuordnung hängt einerseits von der Dominanz eines Topics innerhalb des Dokuments ab und andererseits von der Wahrscheinlichkeit eines Wortes, zu einem bestimmten Topic zu gehören (Steyvers & Griffiths 2007: 7f.). Der Prozess wird 2000-mal repetiert und mit jeder Iteration verändert sich die Berechnung der Dokumentaufteilung bzw. der Topics, damit die Fehler der Rekonstruktion von Worthäufigkeiten in den Original-

dokumenten sinken. Ziel ist eine Balance der Homogenität zwischen Dokumenten – alle Wörter eines Dokuments werden dem gleichen Topic zugeordnet – und Topics – einige Wörter eines Topics haben eine hohe Wahrscheinlichkeit. Die Auflösung der widersprüchlichen Kriterien produziert für jedes Topic Schlüsselbegriffe, die häufig kookkurrent auftreten und damit eine semantische Interpretierbarkeit aufweisen. Aus der Anwendung von LDA geht schliesslich eine Topic-Struktur hervor, die gemäss den Modellannahmen das vorliegende Textkorpus statistisch am besten beschreibt. Topic Modeling mit LDA kombiniert somit eine induktive, *struktursuchende* Vorgehensweise (Baumer et al. 2017) mit statistischen Messverfahren und begünstigt insbesondere explorative Analysen. Das Verfahren eignet sich folglich gut, um ein Wissensgebiet, dessen Konturen noch nicht endgültig umrissen sind, in seiner Heterogenität und Vielschichtigkeit zu erforschen.

Ich berechne die Topic-Modelle sprachgetrennt, weil gemeinsame Modelle aufgrund der Ungleichverteilung (~68 % englisch, ~30 % deutsch) dazu tendieren, die deutschsprachigen Begriffe in einzelnen separaten Topics zu repräsentieren. Dies steht allerdings der Absicht entgegen, die semantischen Strukturen der Dokumente sprachunabhängig zu analysieren. Eine separate Analyse bietet zudem die Möglichkeit, den Einfluss des ökonomischen Feldes auf das Modell zu berücksichtigen, da im englischsprachigen Sample Pharmaunternehmen sehr häufig sind, während sie im deutschsprachigen Sample kaum auftreten.

Die Modellierung der Topic-Struktur mithilfe der LDA erfolgt mit dem Softwarepaket *Topicmodels* in R (Grün & Hornik 2011). Um ein geeignetes Topic-Modell zu bestimmen, habe ich Dutzende von Durchläufen mit jeweils unterschiedlicher Anzahl Topics berechnet und dabei auch quantitative Metriken berücksichtigt. Ein datenbasiertes Vorgehen nimmt meist die *Perplexity-* und *Log-Likelihood*-Werte eines Topic-Modells als Grundlage für die Festlegung der Anzahl Topics (Grün & Hornik 2011: 7; Maier et al. 2018: 99).[20] Die *Perplexity* ist eine Metrik, um die statistische Güte (*goodness-of-fit*) eines Modells zu bestimmen. Dabei wird geschätzt, wie gut ein Modell, das für einen grösseren Teil des Korpus berechnet wurde, einen kleineren Teil der Dokumente vorhersagt (Maier et al. 2018: 99). Die logarithmierte *Likelihood* bewertet, »wie gut ein Set von Parametern (Koeffizienten) geeignet ist, um ein Modell (z. B. eine logistische Regressionsfunktion) an vorhandene Daten anzupassen« (Diaz-Bone & Weischer 2015: 245). Ein höherer *Likelihood*-Wert korrespondiert dabei mit tieferen *Perplexity*-Werten (Maier et al. 2018: 102).

Für das englischsprachige Korpus ergibt die *Perplexity* eine optimale Anzahl bei K = 100 Topics, für die *Log-Likelihood* bei K = 75 Topics. Beim deutschsprachigen Sample liegen die entsprechenden Werte bei K = 90 sowie K = 40 Topics. Da die Struktur von Modellen mit dieser Granularität nicht sinnvoll interpretierbar ist, reduziere ich in nachfolgenden Berechnungen die Anzahl Topics, bis die Modelle inhaltlich valide beschrieben werden können. Dazu dürfen sie nicht zu weit gefasst sein, da sich sonst viele Überschneidungen von Topics ergeben, aber auch nicht zu eng, da sonst viele inhalt-

20 Das R Package LDAtuning umfasst vier unterschiedliche Metriken bzw. Informationskriterien (»Griffiths2004«, »CaoJuan2009«, »Arun2010«, »Deveaud2014«), die für die Festlegung der Anzahl Topics herangezogen werden können. Für das englischsprachige Korpus ergaben die Metriken eine optimale Anzahl Topics zwischen K = 40 und K = 100, für das deutschsprachige Sample lagen die entsprechende Werte zwischen K = 10 und K = 40.

liche Aspekte zusammengefasst werden (Maier et al. 2018: 98). Für die semantische Interpretation der Topics ziehe ich sowohl die statistischen Auftretenswahrscheinlichkeiten der häufigsten Begriffe als auch Dokumente, die durch die jeweiligen Topics gut beschrieben werden, heran. Die resultierenden Modelle, bei denen die häufigsten Begriffe der Topics inhaltlich valide beschrieben werden können, umfassen K = 20 Topics für das englischsprachige und K = 11 Topics für das deutschsprachige Korpus (vgl. Tabellen 5 und 6).

Wie die Festlegung der Anzahl Topics zeigt, spielen qualitative Interpretationsleistungen trotz der Quantifizierung der Textdokumente bzw. ihrer einzelnen Bestandteile eine zentrale Rolle. Die Interpretation der Topics basiert zwar teilweise auf quantitativen Kennzahlen, lässt sich jedoch als solche nicht berechnen (Roberts et al. 2019: 9ff.). Im Sinne eines relationalen Verfahrens erhalten die Topics ihre Sinnhaftigkeit nur im Verhältnis zu anderen Topics und werden erst dadurch interpretierbar. Insofern äussert sich darin eine Anschlussfähigkeit von Topic Modeling an kultursoziologische Ansätze, die die Relationalität von Sinnstrukturen untersuchen (Kirchner & Mohr 2010; Mohr 2000, 2013; Mohr & Bogdanov 2013).

4.3 Strategiedokumente im Feld der Politik

Akteur*innen im Feld der Politik prägen die Herausbildung des Gegenstandes Datenwissenschaften primär durch unterschiedliche Strategiepapiere: Sie entwerfen im Rahmen des Diskurses über die Digitalisierung Szenarien einer vielversprechenden Zukunft und verknüpfen diese mit bildungs- und forschungspolitischen Fördermassnahmen. Politische Strategiepapiere können als kollektive Stellungnahmen und Kompromisse verschiedener, koexistierender ›Flügel‹, Überzeugungen und Weltanschauungen von organisationalen Akteur*innen als sozialen Feldern betrachtet werden. Einerseits sind sie »Resultat einer sozialen Praxis, die von positionsabhängigen Akteursstrategien geprägt ist und die sich wiederum als Teil eines Ensembles von Strategien verstehen lässt, die als Ganzes von einem feldspezifischen Machtkonflikt um Positionen, Ressourcen und Anerkennung getrieben sind, der sich um die Definition von politischen Kategorien dreht« (Bernhard & Schmidt-Wellenburg 2012: 48f.). Andererseits eröffnen die Stellungnahmen, die durch Vielstimmigkeit und Multiperspektivität gekennzeichnet sind, auch Möglichkeiten zur Kooperation mit anderen Akteur*innen. Sowohl durch stärker konflikthafte als auch durch kooperative Praktiken tragen sie zur Konstitution und Permanenz der Datenwissenschaften als neues Wissensgebiet bei.

Ich rekonstruiere die Zukunftsentwürfe durch eine qualitative Inhaltsanalyse und stelle die Verknüpfung der hochschulpolitischen Stellungnahmen zu anderen Feldern, insbesondere der Wissenschaft und der Ökonomie her. Zunächst lege ich dar, wie ich das Korpus der politischen Strategiepapiere zusammengestellt habe (Kap. 4.3.1). Anschliessend erläutere ich das inhaltsanalytische Vorgehen und den Prozess der Kodierung des Materials (Kap. 4.3.2).

4.3.1 Das Korpus der Dokumente

Das Sample umfasst Dokumente und Strategiepapiere von Akteur*innen im Feld der schweizerischen Hochschul- und Forschungspolitik, die sich mit der Digitalisierung in Bildung und Forschung im weiteren Sinne und mit der Herausbildung der Datenwissenschaften als neuem Wissensgebiet im engeren Sinne beschäftigen. Ausgangspunkt für die Zusammenstellung des Samples waren die Dokumente zur Strategie »Digitale Schweiz« des Bundesrates. Über diese konnte ich 34 Strategiepapiere und weitere Dokumente erschliessen, die sich ebenfalls ganz oder teilweise dem genannten Themenfeld widmen. Zu den Akteur*innen gehören sowohl politische Institutionen wie der Bundesrat, das Staatssekretariat für Bildung, Forschung und Innovation (SBFI), das Staatssekretariat für Wirtschaft (SECO), das Bundesamt für Kommunikation (BAKOM), der ETH-Rat, die Konferenz der kantonalen Erziehungsdirektorinnen und -direktoren (EDK) sowie wissenschaftliche Kommissionen, Unternehmensverbände und Think Tanks (vgl. die Liste der untersuchten Dokumente in Tabelle 11 im Anhang).[21]

Der Zeitraum der untersuchten Dokumente reicht von 1998 bis 2020, wobei die überwiegende Mehrheit davon nach 2014 publiziert wurde. Da die Dokumente ab 2014 vermehrt »Big Data« und »Data Science« sowie deren rechtliche, ökonomische und edukative Aspekte in der Bildungs- und Forschungspolitik adressieren, bildet dieser Zeitraum den Schwerpunkt der Analyse. Die älteren Strategiedokumente zur Informationsgesellschaft Schweiz wurden berücksichtigt, um Kontinuitäten und Brüche in den bundesrätlichen Strategien zu untersuchen.

Einen zentralen Bestandteil des politischen Diskurses zur Digitalisierung in der Schweiz bilden die bundesrätlichen Strategiedokumente, Aktionspläne und Berichte zur »Informationsgesellschaft Schweiz« bzw. zur »Digitale[n] Schweiz« (vgl. Tabelle 12 im Anhang): Ende der 1990er-Jahre gab sich der Bundesrat zum ersten Mal eine Strategie für den Umgang mit den »neuen Informations- und Kommunikationstechniken«. Die »Strategie Informationsgesellschaft« wurde im Februar 1998 (im Folgenden: SIG 1998) verabschiedet und 2006 (im Folgenden: SIG 2006) sowie 2012 (im Folgenden: SIG 2012) überarbeitet, wobei Ziele, Grundsätze und Politikbereiche fortlaufend adaptiert wurden (Abun-Nasr 2009; Ciarla 2018). Die im April 2016 lancierte Strategie »Digitale Schweiz« (im Folgenden: SDS 2016) des Bundesrates löst die vorherige »Informationsgesellschaft Schweiz« ab, die im September 2018 (im Folgenden: SDS 2018) revidiert wurde.

Die fünf Strategiedokumente der »Informationsgesellschaft Schweiz« und »Digitale Schweiz« bilden ein Subsample innerhalb des Korpus. Neben ihrer gemeinsamen Autorschaft (Bundesrat) formulieren sie politische Strategien über das Verhältnis von Technologie und Gesellschaft und eignen sich daher besonders für eine vertiefende inhaltliche Analyse über den betrachteten Zeitraum hinweg. Während andere Dokumente im Korpus in erster Linie bildungs- und forschungspolitische Akteur*innen adressieren, richten sich die bundesrätlichen Strategien an ein breiteres, hybrides Publikum.

21 Die untersuchten Dokumente, das Kodierschema sowie das Kodebuch sind auf dem Datenrepositorium FORSbase hinterlegt (Saner & Mützel 2021b).

Formal zu beobachten ist neben den kürzer werdenden Erscheinungszyklen die kontinuierliche Ausweitung des inhaltlichen Umfangs der bundesrätlichen Strategiepapiere. Die Strategie »Digitale Schweiz« 2018 ist fünfmal umfangreicher als die erste Strategie »Informationsgesellschaft Schweiz« von 1998. Der Vergleich bestimmter Wortfrequenzen muss deshalb in Relation zum stark anwachsenden Textumfang der Strategiedokumente gesetzt werden (vgl. Tabelle 12 im Anhang).

4.3.2 Das inhaltsanalytische Vorgehen

Alle Dokumente lagen als Textdateien vor und wurden in der Analysesoftware ATLAS.ti erfasst. Ich habe ausschliesslich jene Textstellen kodiert, die einen expliziten Bezug zu den Bereichen Bildung und Forschung haben. Daraus resultierte ein reduziertes Textsample von rund 350 A4-Seiten. Ich habe die für die Fragestellung relevanten Textstellen in einem offenen, induktiven Kodierungsprozess (Flick 2016: 388ff.) mit zusammenfassenden bzw. erklärenden Kodes versehen (Friese 2012: 92ff.). Dabei entstand in einem ersten Durchgang ein Kategoriensystem, das ich in insgesamt drei Durchgängen überprüft und überarbeitet habe. Durch das mehrmalige Lesen und Kodieren der Textstellen konnte ich das Kategoriensystem erweitern und ausdifferenzieren (vgl. das Kodierschema in Saner & Mützel 2021b). Ziel war dabei neben der inhaltlichen Strukturierung des Materials das Eruieren der relevanten Themen und Dimensionen im Korpus der politischen Dokumente.

Die Kodierung der Texte in drei Durchläufen ergab 75 unterschiedliche Kodes, die insgesamt 1144 Zitate umfassen.[22] Bei Passagen, die mit mehr als einem Kode versehen sind, handelt es sich um inhaltlich dichte Textstellen mit unterschiedlichen thematischen Bezügen. Ein Beispiel für ein inhaltlich äusserst dichtes Zitat lautet wie folgt:

> »Um die mit dem Strukturwandel verbundenen Chancen zu nutzen und Herausforderungen erfolgreich bewältigen zu können, müssen diese bereichsübergreifend sowohl national als auch international vernetzt angegangen werden.« (SDS 2018: 4)

Der erste Teilsatz bezieht sich auf den »Strukturwandel«, der auf die Digitalisierung als gesellschaftlicher Transformationsprozess (1. Kode) verweist, wodurch sich sowohl »Chancen« (2. Kode) als auch »Herausforderungen« (3. Kode) eröffnen – dies bildet, wie ich noch zeigen werde, ein zentrales Element im politischen Diskurs über Digitalisierung von Bildung und Forschung. Der zweite Teilsatz bezieht sich auf die Strategien, um die »Chancen zu nutzen und Herausforderungen erfolgreich zu bewältigen«, wobei hier Interdisziplinarität (4. Kode) sowie Kooperation und Vernetzung (5. Kode) erwähnt werden.

Die Kodierung der Zitate erfolgte in erster Linie über bestimmte Begriffe. Beispielsweise umfassen die Textstellen, die inhaltlich mit »Digitalisierung als gesellschaftlicher Transformationsprozess« kodiert werden, sehr oft prozessindizierende Begriffe wie Automatisierung, Modernisierung, (Struktur-)Wandel, Veränderung oder Entwicklung. Mittels einfacher Häufigkeitszählung können diese Begriffe in den einzelnen Dokumenten aggregiert ausgewertet werden. Dies erlaubt es, Entwick-

22 Zwei Drittel der zitierten Stellen sind einmalig kodiert (66.7 %), die restlichen Zitate sind zweimalig (25.7 %), dreimalig (6.6 %) oder häufiger (1 %) kodiert.

lungen bestimmter zentraler Kategorien im politischen Diskurs über den Zeitverlauf hinweg zu analysieren. Dies ist insbesondere für das Subkorpus der fünf Strategiedokumente interessant, da diese über zwei Jahrzehnte hinweg soziotechnische Zukunftsszenarien für die Schweiz formulieren.

Zudem lassen sich die identifizierten Kodes in neun übergreifenden Dimensionen zusammenfassen (vgl. Tabelle 8 im Anhang). Am häufigsten sind Zitate in der politischen Dimension zu verorten, was auf die Verortung der Akteur*innen im politischen Feld verweist, gefolgt von der wissenschaftlichen Dimension, was die thematische Orientierung der untersuchten Dokumente indiziert. An dritter Stelle folgt die Datendimension, die das Erhebungskriterium für die Auswahl des Korpus darstellt. Ferner sind auch die organisationale, technische, ökonomische sowie edukative Dimension in den Dokumenten bedeutend. Etwas weniger prävalent sind die internationale Dimension sowie die Kompetenzen, die aber dennoch in der grossen Mehrheit der Dokumente enthalten sind.

Knapp die Hälfte der Kodes (46 %) bildet eine Dimension ab. Die übrigen Kodes sind in zwei (39 %) oder mehr (15 %) Dimensionen enthalten. Die Multidimensionalität vieler Kodes bildet das relationale Moment der Analyse, indem Kodes zwei oder mehrere Dimensionen miteinander verknüpfen. So kann der Kode »Standortwettbewerb« sowohl in der ökonomischen, wissenschaftlichen als auch internationalen Dimension massgebend sein. Um dies anhand eines Zitats aus der aktuellen Strategie »Digitale Schweiz« zu veranschaulichen:

> »Um den Spitzenplatz der Schweiz als Innovations- und Forschungsstandort zu halten, müssen die Forschungskompetenzen bezüglich digitaler Technologien in ihrer ganzen Breite gestärkt und der Wissenstransfer in die Wirtschaft beschleunigt werden.« (SDS 2018: 5)

Über die internationale Dimension, den »Spitzenplatz der Schweiz« in einem nicht weiter bezeichneten Vergleich, wird die wissenschaftliche mit der ökonomischen Dimension verknüpft (»Innovations- und Forschungsstandort«, »Wissenstransfer in die Wirtschaft beschleunigen«). Durch die Einbettung und den Kontext der identifizierten Textstellen sowie die Beziehungen zwischen den verschiedenen Dimensionen ergibt sich damit die Bedeutung bzw. Sinnhaftigkeit eines Konzeptes wie »Standortwettbewerb«.

4.4 Curricula und Interviews im akademischen Feld

Die Curricula aller verfügbaren Studienangebote in Data Science an Schweizer Universitäten und Hochschulen sowie Interviews mit Lehrenden bilden das empirische Material für das akademische Feld, das ich mittels qualitativer Inhaltsanalysen untersuche. Während die Curricula im Falle der Datenwissenschaften meist Produkte grenzüberschreitender disziplinärer Kooperation darstellen, treten in den Interviews die epistemologischen und disziplinären Perspektiven auf den Gegenstand deutlicher zutage. Die Kombination der Materialien erlaubt es, die Widersprüche und Dynamiken synchroner Praktiken von Grenzziehung und Grenzüberschreitung

in der Herausbildung der Datenwissenschaften als einem Möglichkeitsraum zu adressieren.

Zunächst beschreibe ich das Sample, das die Curricula aller Studiengänge bzw. Vertiefungen in Datenwissenschaften an Schweizer Universitäten und Hochschulen umfasst. Dieses kodiere und untersuche ich mithilfe der Inhaltsanalyse (Kap. 4.4.1). Anschliessend skizziere ich den Prozess der Erhebung der qualitativen Interviews mit den Studiengangleitenden ausgewählter Studiengänge (Kap. 4.4.2), die ich ebenfalls inhaltsanalytisch auswerte (Kap. 4.4.3).

4.4.1 Curricula und Curricula-Analyse

Wie bereits im Forschungsstand dargelegt, sind Curricula mehr als die Summe der relevanten Wissensbestände eines Feldes: In ihnen artikulieren sich stets auch Konzeptionen über das Verhältnis eines Wissensgebiets zu seinen ökonomischen, technologischen, politischen und weiteren Bezugsfeldern. Curricula der Datenwissenschaften können als Kompromissprodukte verschiedener Disziplinen verstanden werden, die Aushandlungen der involvierten Disziplinen innerhalb von Universitäten oder Hochschulen als organisationalen Feldern abbilden (Knight et al. 2013). Insofern repräsentieren Curricula – ähnlich wie Stellenanzeigen – durch die Kombination von Strukturen und Inhalten eine bestimmte Deutung des Wissensgebiets Datenwissenschaften. Durch die Vielzahl konkurrierender Curricula resultieren und koexistieren demnach multiple, vielstimmige Perspektiven.

Um die verschiedenen curricularen Deutungen der Datenwissenschaften analysieren zu können, habe ich ein Sample aller Studienangebote in Data Science an Schweizer Universitäten, ETH und Fachhochschulen zusammengestellt. Bezeichnend für das Material ist, dass das Sample kontinuierlich angewachsen ist: Gab es zum Zeitpunkt der ersten Erhebungsphase (Sommer 2017) lediglich neun Studiengänge oder Vertiefungen, so ist die Zahl bis zum jetzigen Zeitpunkt (Februar 2022) auf 42 Angebote in unterschiedlichen Hochschultypen und Studienniveaus angewachsen (vgl. Tabelle 1). Die Entwicklung meines Sample bildet insofern den Diffusionsprozess einer Grenzkategorie ab, die von einer wissenschaftlich-industriellen Schnittstelle zu einem disziplinen- und feldübergreifenden Wissensgebiet mutierte, dem hohe Legitimität zugesprochen wird.

Tabelle 1: Übersicht der Studiengänge und Interviews nach institutionellem Kontext[23]

Hochschultyp	Studienstufe	Anzahl Studiengänge (davon Vertiefungen)	Interviews durchgeführt
ETH	Master	2	3
Universität	Master	11 (3)	3
	Bachelor	1 (1)*	1
ETH & Universität	Weiterbildung (CAS/DAS/MAS)	7***	1
Universität & Fachhochschule	PhD-Programm[24]	1	0
Fachhochschule	Master	3 (1)	1
	Bachelor	6 (3)	0**
	Weiterbildung (CAS/DAS/MAS)	12	2
Forschungsinstitute (SDSC, Datalab)	Weiterbildung (CAS/DAS/MAS)	1***	2
Total		42	13

* es handelt sich um ein Lehrprogramm mit Zertifikat (ohne Abschluss BA/MA)

** die Studiengänge wurden erst nach der qualitativen Erhebungsphase etabliert

*** ein Programm wird in Kooperation von Universität und ETH Lausanne (SDSC) durchgeführt

Mittels Websuche erstellte ich zunächst eine Liste aller verfügbaren Studienangebote in »Data Science«. Da einige Studienangebote lediglich als Nebenfach (Minor), Vertiefungen oder Profile innerhalb anderer Studiengänge (vor allem Informatik) angeboten werden, schloss ich fortan auch solche Programme ein. Parallel öffnete ich den Fokus und inkludierte nun auch Studiengänge, die Data Analytics oder Data Management in der Bezeichnung tragen. Zudem wurden ab 2017 erste Studienangebote in Artificial Intelligence oder Machine Learning etabliert, die ebenfalls die Verarbeitung und Analyse grosser Datenmengen umfassen (vgl. Tabelle 13 im Anhang).

Nachdem die Suchstrategie nun differenziert und verfeinert war, sammelte ich die für die Studienangebote verfügbaren öffentlichen Dokumente und Webseiten wie Werbe- und Informationsmaterialien, Vorlesungsverzeichnisse und Kursbeschreibungen. Ich begann hochschulübergreifende Themen zu eruieren, die die verschiede-

23 Stand: Februar 2022.

24 Das PhD Network in Data Science der Universitäten Zürich und Neuenburg in Kooperation mit den Fachhochschulen ZHAW und SUPSI wird von *Swissuniversities* im Rahmen des Teilprojekts 2 »Kooperation zwischen Schweizer FH/PH und UH« gefördert. Vgl. https://www.swissuniversities.ch/themen/nachwuchsfoerderung/p-1-doktoratsprogramme/tp2-kooperation-zwischen-fh/ph-und-uh (Zugriff: 03.02.2022).

nen Studiengänge und Vertiefungen charakterisieren. Anschliessend kodierte ich die Materialien inhaltsanalytisch nach unterschiedlichen Merkmalen (Tabelle 2).

Tabelle 2: Merkmale und Ausprägungen der Studiengänge im Sample

Merkmal	**Ausprägungen**
Hochschultyp	Universität, ETH, Fachhochschule
Name	Name Programm/Vertiefung
Studienstufe	Bachelor, Master, PhD, Weiterbildung
Organisationale Verortung	Fakultäten bzw. Departemente
Kooperation intern	Fakultäten bzw. Departemente
Kooperation extern	Universität bzw. Hochschule
Fachliche Verortung der Lehrenden	Denomination der Professuren*
Zeitpunkt der Etablierung	Jahr
Zulassungsvoraussetzungen	Unterschiedliche Kriterien (Bachelorabschlüsse, Lehrveranstaltungen etc.)
Umfang	Anzahl ECTS
Pflichtbereich	ja/nein, Anzahl ECTS
Wahlbereich	ja/nein, Anzahl ECTS
Abschlussarbeit	ja/nein, Anzahl ECTS
Praxisorientierte Kurse	ja/nein, Anzahl ECTS
Weitere Inhalte	ja/nein, Anzahl ECTS
Publikum	Zielgruppen, Rekrutierungsbemühungen

* Datenlage unvollständig

Die erhobenen Merkmale erfassen demnach sowohl die institutionelle Verankerung, die Struktur und den Aufbau sowie inhaltliche Aspekte der Studiengänge im Sample. Der Hochschultyp markiert die Verortung im akademischen Feld der Schweiz, die nicht nur mit der Grösse und historischen Traditionen, sondern auch mit politischen Verantwortlichkeiten, Vorgaben und finanziellen Ressourcen korrespondiert. Dazu gehört, wie ich noch zeigen werde, auch der Etablierungszeitpunkt, der auf unterschiedliche Strategien und Positionierungen, aber auch organisationale Möglichkeiten zur Einrichtung neuer Studienangebote verweist. Merkmale wie die Programmbezeichnung, Studienstufe, Verortung sowie interne bzw. externe Kooperationen verweisen einerseits auf die organisationale Eingliederung, andererseits auch auf die inhaltliche Ausrichtung des Studiengangs. Die Analyse des Aufbaus der Studiengänge fördert eine spezifische Struktur zutage, die sich einerseits an ingenieurwissenschaftliche Traditionen anlehnt, diese gleichzeitig aber auch erweitert (vgl. Kap. 9): So können die curricularen Elemente – in der Regel Module genannt – von Kernbereich,

Wahlbereich, Abschlussarbeit, praxisorientierte Kurse sowie weitere (»komplementäre«) Inhalte differenziert werden, die für die vorliegende Analyse als Merkmale festgelegt werden. Über die fachliche Verortung der Lehrenden sowie die jeweilige Anzahl Credits in den einzelnen Bereichen erschliessen sich die disziplinären Verankerungen bzw. die eigentliche Interdisziplinarität des Lehrangebots der Studiengänge.

In Anlehnung an Knight et al. (2013: 147) unterscheide ich die Regelstudiengänge (auf Bachelor- und Masterstufe) im Sample in ein Kontinuum zwischen starken und schwachen interdisziplinären Programmen anhand der Zusammensetzung sowie der Anzahl verpflichtender Credits im Kernbereich: Setzt sich der Kernbereich aus mehreren disziplinären Veranstaltungen zusammen und liegt der Anteil verpflichtender Kreditpunkte über 50 %, handelt es sich in diesem Sinne um starke interdisziplinäre Programme.[25]

4.4.2 Interviews mit Studiengangleitenden und Professor*innen

In einem zweiten Erhebungsschritt habe ich qualitative Interviews mit Leitungspersonen und Lehrenden der Studiengänge durchgeführt.[26] Die Interviews dienen dazu, die strukturellen und inhaltlichen Merkmale der Curricula-Analyse mit den Verantwortlichen der Studiengänge zu diskutieren. Einerseits trägt dies dazu bei, die Erkenntnisse des ersten Analyseschritts besser zu verstehen und mit Expert*innen aus dem Untersuchungsfeld zu verifizieren. Andererseits konnte ich so weitere Erkenntnisse, etwa über strategische Überlegungen und organisationale Herausforderungen, gewinnen, die durch die institutionelle Verortung und Interdisziplinarität der Curricula induziert sind und sich nicht aus den öffentlich zugänglichen Dokumenten eruieren lassen. Die Kombination von Curricula- und Interviewanalysen erschliesst somit jene Konstruktionsleistungen, durch welche Akteur*innen im akademischen Feld die Datenwissenschaften objektiv wie subjektiv rahmen.

Insgesamt habe ich dreizehn semistandardisierte Interviews mit vierzehn Studiengangleitenden, Professor*innen und Dozierenden durchgeführt, die ich als Expert*innen für »Data Science« adressiert habe. Die Interviews, die zwischen dreissig Minuten und zwei Stunden dauerten, habe ich auf Tonband aufgezeichnet.[27] Unmittelbar im Nachgang an die Interviews habe ich jeweils Memos verfasst, in denen ich neben Notizen und ersten wichtigen Erkenntnissen des Gesprächs auch meine subjektiven Eindrücke über die Interviewsituation festhielt. Zudem sind auch Reflexionen über meine Rolle als Interviewer enthalten. Memos unterstützen nicht nur den

25 Knight et al. schlagen als zweites, organisationales Merkmal den prozentualen Anteil jener Lehrenden vor, deren Denomination innerhalb des Programms liegt, und ob die Studiengangleitung ihre Professur innerhalb oder ausserhalb des Programms hat. Da allerdings im akademischen Feld der Schweiz bis dato kaum Professuren in Data Science existieren, liefert die Erhebung dieses Merkmals keinen zusätzlichen Erkenntnisgewinn.

26 Das ursprüngliche Vorhaben, eine Erhebung bei allen Studiengängen im Feld durchzuführen, liess sich aufgrund mangelnder Teilnahmebereitschaft leider nicht realisieren. Deshalb habe ich das Ziel verfolgt, mit unterschiedlichen Positionen im Bereich »Data Science« im universitären Feld der Schweiz zu sprechen.

27 Ein Interview habe ich via E-Mail geführt, wodurch sich die Möglichkeit, Nach- und Vertiefungsfragen zu stellen, erheblich reduziert hat.

späteren Analyseprozess, indem sie Anleitungen geben für die Relevanzsetzungen in den Interviews, sondern sind auch ein wichtiges Hilfsinstrument einer reflexiven Wissensproduktion (Kühner et al. 2013). Die Interviews habe ich in vollständiger Länge nach vorgängig definierten Regeln transkribiert (Langer 2010).[28]

Eingeleitet habe ich die Interviews jeweils mit der Frage nach den wissenschaftlichen und berufsbiographischen Werdegängen der Befragten (vgl. den Interviewleitfaden im Anhang). Nimmt man die Antworten sowie die gegenwärtigen Stellenbezeichnungen als Ausgangspunkt für eine fachliche Verortung der Interviewpartner*innen, so präsentiert sich angesichts der postulierten Heterogenität des Wissensfeldes ein eher einseitiges Bild: Gesprächspartner*innen waren neun Informatiker*innen sowie je eine Person mit disziplinärem Hintergrund in Chemie, Mathematik, Ökonomie, Informationssysteme sowie Marketing und Kommunikation. Obwohl diverse Befragte teilweise lange Berufsbiographien in Datenwissenschaften aufweisen, verwendet niemand die Selbstbeschreibung als »Data Scientist« für das eigene Tätigkeitsgebiet. Zum einen sind die Lehrenden in bestimmten Disziplinen fachkulturell sozialisiert (Holley 2009), sodass die Neuheit des interdisziplinären Gegenstandes nachgelagert ist: Die Offenheit des Raumes zwischen etablierten Disziplinen bietet die Möglichkeit, über Forschung, Lehre und anderweitig darin tätig zu sein, ohne sich selbst damit identifizieren zu müssen. Zum anderen manifestieren sich darin auch innerakademische Distinktionsmechanismen (Bourdieu 1988), die »Data Science« als professionelle Praxis ausserhalb des akademischen Feldes situieren und primär kommerziellen Tätigkeitsbereichen in der »Industrie« zuschreiben. Schliesslich existieren, obwohl »Data Science« als Wissens- und Forschungsgebiet anerkannt und explizit in Stelleninseraten gesucht wird, trotz der zahlreichen Studiengänge und Vertiefungen an den untersuchten akademischen Einrichtungen bis dato nur einzelne Professuren mit Denomination »Data Science«.[29]

Der Interviewleitfaden beinhaltet Fragen zu den folgenden Themengebieten: Aufbau und Curriculum des Studiengangs, Kursinhalte, Studienplätze und Zusammensetzung der Studierenden, Relationen zu anderen Hochschulen, Arbeitsmarkt, Definitionen von »Data Science« sowie Hinweise für die weitere Forschung (vgl. den Interviewleitfaden im Anhang). Den Leitfaden habe ich je nach Interview flexibel und situativ eingesetzt, wobei ich den Relevanzsetzungen der Interviewten, die sich im Gespräch ergaben, und den sich daraus ergebenden Anschlussfragen hohes Gewicht beigemessen habe.

Zur Anonymisierung der Aussagen der Befragten unterscheide ich nicht zwischen Studiengangleitenden, Professor*innen und Dozierenden, sondern verwende eine Abkürzung für alle Interviewten in Kombination mit dem Hochschultypus (ETH, UH, FH), einem individuellen Buchstaben sowie der Interview- und Zitatnummer, um eine Textstelle eindeutig einem Interview zuzuordnen. So steht die Abkürzung »Prof_ETH_A« für Professor*innen an den beiden ETH, »Prof_UH_B« für Interviewte

28 Die Interviewdateien, Transkripte und der Abkürzungsschlüssel zur Identifizierung der Befragten sind in einem persönlichen, passwortgeschützten Ordner auf dem akademischen Server SWITCHdrive gespeichert. Dieses Vorgehen schützt die Persönlichkeitsrechte der Interviewten und garantiert den Datenschutz.

29 Zudem werden an akademischen Einrichtungen kaum Positionen für »Data Scientists« eingerichtet (Geiger et al. 2018; Moore-Sloan Data Science Environments 2018).

an Universitäten und »Prof_FH_C« für Gesprächspartner*innen an Fachhochschulen. Ist eine Identifizierung einer Universität oder Hochschule notwendig für den Kontext einer Aussage, wird die Nennung der interviewten Person sowie der Interview- und Zitatnummer weggelassen.

4.4.3 Qualitative Analyse der Interviews

Zur Unterstützung der qualitativen Analyse habe ich die Transkripte in der Analysesoftware ATLAS.ti erfasst. In einem ersten Schritt habe ich die für das Erkenntnisinteresse und die Fragestellungen relevanten Textstellen mit zusammenfassenden bzw. erklärenden Kodes versehen (Friese 2012: 92ff.). Bei der Auswertung gemäss der strukturierenden qualitativen Inhaltsanalyse wird eine inhaltlich-typisierende Strukturierung vorgenommen und der Umfang des Materials reduziert (Mayring 2010: 92). In die Erarbeitung des Kategoriensystems flossen neben dem Forschungsstand zu interdisziplinären Curricula (und insbesondere jene der Datenwissenschaften) auch die Ergebnisse über die strukturelle und inhaltliche Verfasstheit der untersuchten Curricula mit ein. Ferner habe ich im Sinne eines offenen, induktiven Kodierungsprozesses (Flick 2016: 388ff.) auch Kodes aus den Interviews selbst gewonnen. Dabei entstand ein integrales Kategoriensystem, das ich nach dreimaligem Kodieren derselben Dokumente kontinuierlich erweitern und ausdifferenzieren konnte.

Grundlegendes Ziel dabei war das Eruieren der relevanten Aussagen, Einschätzungen und Reflexionen, die Lehrende über das Feld der Datenwissenschaften anstellen. Die Stellungnahmen der Interviewten erweitern und vertiefen insbesondere die strukturellen und inhaltlichen Merkmale, die ich aus der Analyse der Curricula gewinnen konnte. Sie geben mit anderen Worten den beobachtbaren organisationalen Implementierungen und curricularen Eigenheiten einen subjektiven Sinn. Zusätzlich ermöglicht die Bezugnahme auf die disziplinäre Verankerung der Befragten sowie deren Positionen im akademischen Feld eine Kontrastierung verschiedener Aussagen, wodurch sich sowohl disziplinäre als auch feldspezifische Grenzziehungen und -überschreitungen rekonstruieren lassen.

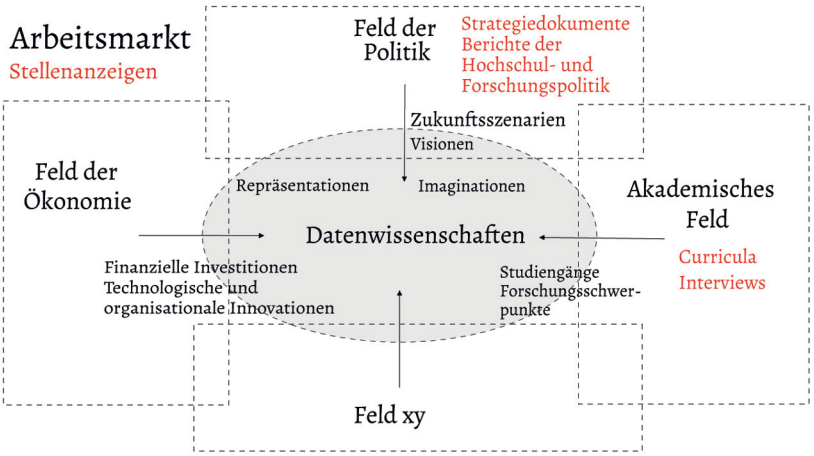
Die qualitative Inhaltsanalyse identifiziert vier zentrale Themenbereiche in den Interviews, die auf unterschiedlichen gesellschaftlichen Ebenen verortet sind und zwischen denen multiple Wechselwirkungen existieren. Zunächst beschreiben und verorten die Befragten die Genese der Datenwissenschaften im Kontext eines fundamentalen Wandels wissenschaftlicher Erkenntnisproduktion. Sie äussern je nach disziplinärer Sozialisation und institutioneller Verankerung unterschiedliche Perspektiven auf den Gegenstand. Die multiplen Verständnisse markieren nicht nur die Heterogenität der Datenwissenschaften, sondern strukturieren wiederum die Positionierung, Ausgestaltung und Profilbildung von Studienprogrammen. Ein zweites Themenfeld umfasst politische und ökonomische Faktoren sowie hochschulstrategische Überlegungen, die den Aufbau konkreter Studienprogramme beeinflussen. Drittens findet das Zusammenspiel hochschulexterner und -interner Einflussgrössen eine Umsetzung in konkreten Implementierungsprozessen der Curricula in Hochschulen und Universitäten. Schliesslich skizzieren die Befragten viertens umfangreiche Kataloge jener als Kompetenzen bezeichneten individuellen Eigenschaften, die sie Praktiker*innen der Datenwissenschaften zuschreiben. Indem die für Datenwissenschaften relevanten Kompetenzen als Zukunftsressourcen gerahmt werden, werden

sie wiederum mit den durch makrostrukturelle Transformationen induzierten neuen Modi der Wissensproduktion verknüpft.

4.5 Zusammenfassung

Die verteilte Analyse von Zwischenräumen bildet ein flexibles, adaptierbares empirisches Modell, das mit Mehrebenenanalysen kompatibel ist: In der vorliegenden Arbeit fokussiere ich primär auf die feldübergreifende Makroebene von Stellenanzeigen sowie die Feldebene der Hochschul- und Forschungspolitik einerseits und auf die verschiedenen Organisationsformen im akademischen Feld andererseits (vgl. Abbildung 2).

Abbildung 2: Methodische Vorgehensweise



Auf der Makroebene ermöglicht die Erhebung von Stellenanzeigen, die multiplen Bedeutungskonstruktionen der Datenwissenschaften durch organisationale Akteur*innen im Arbeitsmarkt zu analysieren. Die deskriptive Analyse erschliesst die zentralen Strukturen des Arbeitsmarktes und zeigt Häufigkeiten und Verteilungen an. Die Auswertung mittels Topic-Modeling-Verfahren identifiziert sodann latente Themen in den Dokumenten. Diese werden nicht ausschliesslich einem Thema alleine zugeordnet, sondern sind in verschiedenen Themen enthalten. Über deren Distribution und Zusammensetzung lässt sich die Vielstimmigkeit sowohl feldtypischer als auch feldübergreifender Themen und Inhalte der Datenwissenschaften in den Stellenanzeigen zugleich rekonstruieren.

Danach wechsle ich die Untersuchungsebene und untersuche Ausprägungen des Gegenstands in zwei zentralen Feldern: Im Feld der Politik entwerfen Akteur*innen durch Strategiepapiere im Rahmen des Diskurses über die Digitalisierung Szenarien einer vielversprechenden Zukunft durch die breite Anwendung der Datenwissenschaften. Durch die inhaltsanalytische Kodierung der Dokumente erfasse ich die artikulierte Multidimensionalität der verwendeten Begriffe und Konzepte und setze sie

in Beziehung zu den anderen Feldern, insbesondere der Wissenschaft sowie der Ökonomie. Den Gegenstand analytisch als verteilt zu konzipieren, erlaubt es somit, die multiplen Bedeutungsdimensionen wechselseitig aufeinander zu beziehen.[30]

Während im Arbeitsmarkt und im Feld der Politik der Fokus der Betrachtung auf Praktiken von Begriffsarbeit liegt, deren feldübergreifende Effekte Zwischenräume als Möglichkeitsräume eröffnen, stehen im akademischen Feld die Wechselwirkungen von Begriffs- und Grenzarbeit im Zentrum: In Curricula und Interviews, d. h. in kollektiven und individuellen Stellungnahmen, formulieren Akteur*innen der Wissenschaft disziplinäre bzw. epistemische Perspektiven auf die Datenwissenschaften. Durch das inhaltsanalytische Vorgehen erfasse ich nicht nur die verbindenden Momente der verwendeten Begriffe, sondern auch die widersprüchlichen Dynamiken und Potenziale, die sich in den synchronen Praktiken von Grenzziehung und Grenzüberschreitung artikulieren.

Eine verteilte Analyse von Räumen zwischen Feldern erschliesst somit die multiplen, mitunter divergierenden Bedeutungskonstruktionen des sozialen Phänomens Datenwissenschaften auf unterschiedlichen Ebenen. Indem der Schwerpunkt der Analyse weniger auf die Strukturen der einzelnen Felder, sondern auf die feldübergreifenden Effekte von kollektiven Praktiken gelegt wird, können die Synchronizitäten und Parallelitäten von Praktiken der Begriffs- und Grenzarbeit in der Fundierung und Konturierung der Datenwissenschaften als verteiltes, zwischenräumliches Phänomen herausgearbeitet werden.

30 Daran kann auch eine mikrosoziologisch fundierte Studie anschliessen, die sich die Sinnkonstruktion in Datenpraktiken in einzelnen Organisationen oder Settings anschaut (Mützel et al. 2018).