Neben strukturierten Daten, die nicht selten explizit für Forschungszwecke erhoben wurden, werden in den Wirtschafts- und Sozialwissenschaften zunehmend unstrukturierte Daten für die Forschung relevant. Diese Daten stammen häufig aus nicht-standardisierten Quellen, wie etwa von Webseiten, aus digitalen Geschäftsberichten, Sozialen Medien usw., sie kommen in vielfältigen Formaten (Audio, Video, Text, Bild oder multimodal) vor und können sehr umfangreich sein. Hieraus resultieren neue Herausforderungen, die das Konsortium BERD@NFDI als Beitrag zum Aufbau der Nationalen Forschungsdateninfrastruktur (NFDI) adressieren wird. Der Beitrag gibt zunächst einen Überblick über die Ausgangslage, die Mission sowie die Zusammensetzung des Konsortiums und zeigt, warum die Verwendung unstrukturierter Daten ein erweitertes Modell der empirischen Forschung bedingt. Nach einer ersten Analyse des Bedarfs der Community wird das Arbeitsprogramm von BERD@NFDI und abschließend die tragende Rolle der an BERD@NFDI beteiligten Bibliotheken dargestellt.

In addition to structured data, which is often collected explicitly for research purposes, unstructured data is increasingly relevant for research in economics and the social sciences. These data often originate from non-standard sources, such as websites, digital business reports, social media, etc., they come in diverse formats (audio, video, text, image or multimodal) and can be very large in scale. This results in new challenges that the BERD@NFDI consortium will address in its contribution to the development of the National Research Data Infrastructure (NFDI). The paper first gives an overview of the starting position, the mission as well as the composition of the consortium. It goes on to show why the use of unstructured data requires an extended model of empirical research. Then it presents a preliminary analysis of the needs of the community, the work programme of BERD@NFDI and, finally, the supporting role of the libraries participating in BERD@NFDI.

SABINE GEHRLEIN, ANNETTE KLEIN, IRENE SCHUMM, KLAUS TOCHTERMANN

BERD@NFDI für unstrukturierte Daten in den Wirtschafts- und Sozialwissenschaften

Ausgangslage, Mission und Konsortium

Die Forschung in den Wirtschafts- und Sozialwissenschaften befasst sich mit den Beziehungen zwischen Individuen und Organisationen in einer Gesellschaft. Um diese komplexen Systeme zu verstehen, wenden diese Disziplinen seit Langem empirische Methoden auf strukturierte Daten an. Ein klassisches Beispiel sind intensiv und aufwendig vorbereitete Umfragen, die entsprechend standardisiert auswert- und darstellbar sind. Zunehmend gewinnen jedoch auch unstrukturierte Daten aus nicht-standardisierten Quellen an Relevanz, d. h. Informationen, die entweder kein zuvor standardisiertes Datenmodell haben oder nicht in einer vordefinierten Weise organisiert sind, z. B. die Darstellung auf Unternehmenswebseiten oder auch Texte, Bilder und Videos aus Sozialen Medien. Die Erzeugung stetiger Datenströme in Gesellschaft und Wirtschaft verstärkt diesen Trend: Schätzungen zufolge werden bis 2025 rund 80 % der in der Wirtschaft verarbeiteten Daten unstrukturierter Natur sein.[1]

Aufgrund des enormen Umfangs, aber vor allem wegen der fehlenden Struktur und der Heterogenität der Rohdaten benötigt die wissenschaftliche Community innovative und nachnutzbare Methoden, insbesondere aus den Bereichen künstliche Intelligenz und maschinelles Lernen, sowie eine geeignete Speicher- und Rechenumgebung, um diese Daten aufbereiten und so für die wissenschaftliche Analyse nutzbar machen zu können. Die Verfahren zur Gewinnung, Bereitstellung, Aufbereitung und Analyse der Daten werden zu einem integralen Bestandteil des Lebenszyklus von Forschungsdaten und müssen folglich genauso gepflegt und bewahrt werden wie die Daten selbst. Die Mission von BERD@NFDI[2] ist es, dieses Desiderat zu erfüllen und für die wirtschafts- und sozialwissenschaftliche Fachcommunity eine zukunftsorientierte und leistungsfähige Infrastruktur für das integrierte Management unstrukturierter Daten und wissenschaftlicher Software zu entwickeln.

Das Konsortium BERD@NFDI unter der Leitung der Universität Mannheim besteht aus sechs Institutionen, darunter Forschungseinrichtungen und Infrastrukturanbieter mit Schwerpunkt auf den Wirtschafts- und Sozialwissenschaften, die in unterschiedlicher Zusammensetzung bereits seit Langem kooperieren. Forschende der Universitäten Mannheim, Hamburg, Köln und LMU München tragen ihre Kompetenzen in den Wirtschafts- und Sozialwissenschaften, Data Science und Machine Learning bei. Die infrastrukturelle Basis

https://doi.org/10.3196/1864295020691280 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY-SA

bilden das Leibniz-Informationszentrum Wirtschaft (ZBW), die Universitätsbibliothek Mannheim sowie das Leibniz-Institut für Sozialwissenschaften (GESIS). Rechen- und Speicherkapazitäten werden durch das Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften (LRZ) sowie die Universitäts-IT Mannheim beigesteuert.

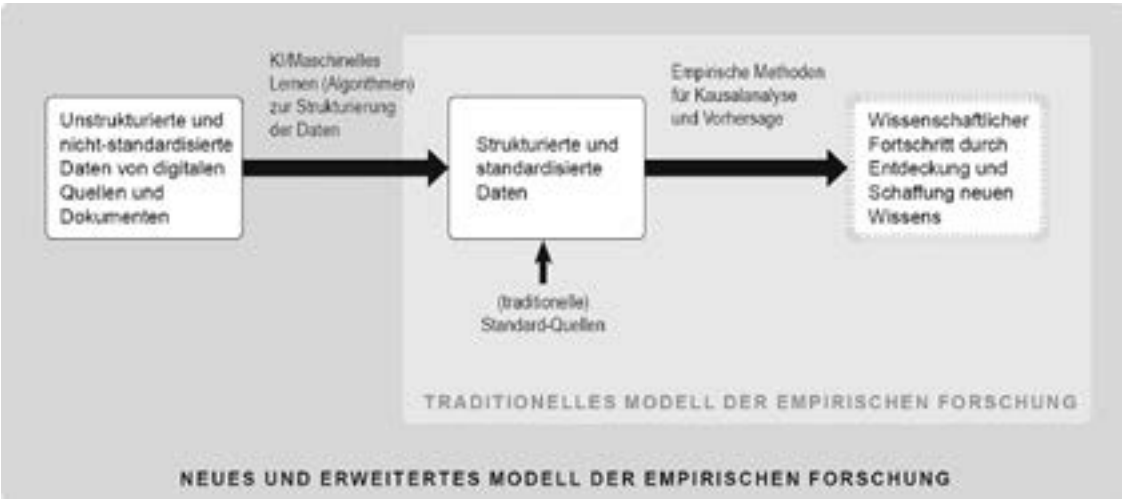
Als Nukleus des Konsortiums kann das Science Data Center BERD@BW (Business and Economic Research Data Center in Baden-Württemberg) gelten. Es entstand im Jahr 2018 zunächst auf Ebene des Landes Baden-Württemberg als Initiative aus Forschenden und Infrastruktureinrichtungen der Universität Mannheim und des Zentrums für Europäische Wirtschaftsforschung (ZEW).[3] Schnell waren deutschlandweit weitere Partner gewonnen, sodass die Nationale Forschungsdateninfrastruktur schließlich die große Chance bot, diese Aktivitäten gemeinsam weiter auszubauen und auf ein nationales Level zu heben.

Erweitertes Modell für unstrukturierte Daten

Traditionell werden in den Sozialwissenschaften Daten, z. B. über Einzelpersonen oder Unternehmen, standardisiert und strukturiert erhoben. Die Datengenerierung, so etwa bei Befragungen, erfolgt in der Regel auf der Basis eines klar durchdachten und vorbereiteten Designs. Entsprechend strukturiert können die erfassten Daten dann im Anschluss relativ einfach ausgewertet, Kausalanalysen durchgeführt und Vorhersagen getroffen werden. Dieses Modell ist als »Traditionelles Modell der empirischen Forschung« zu fassen (s. Abb. 1).

Die traditionelle Herangehensweise in der empirischen Forschung wird seit geraumer Zeit durch ein Modell erweitert und ergänzt, welches auf eine neue Art der Datengenerierung fußt. Im Internet, in den Sozialen Medien und mithilfe von digitalen Technologien, wie z. B. Sensortechnologien, erzeugen die Menschen ständig riesige Mengen an Daten, ohne sich jedoch darüber bewusst zu sein oder gar die Zielsetzung zu haben, damit Forschungszwecken zu dienen. Dieser Prozess der stetigen Datengenerierung wird in der Literatur als Datafication oder Datafizierung bezeichnet und drückt aus, dass viele Aspekte des täglichen Lebens Datenspuren hinterlassen.[4] So etwa schaffen Streaming-Dienste, wie Spotify, Amazon Prime Music oder Youtube, eine noch nicht dagewesene Fülle an Daten, um die Vorlieben und Gewohnheiten ihrer Hörer*innen zu verstehen, die wiederum auch für die Forschung von Bedeutung sind: Wer hört welche Musik, wie oft, zu welcher Zeit, welche Musik wird geteilt, worüber sprechen die Hörer*innen, welche Bilder posten sie, etc. Unstrukturierte und nicht-standardisierte Daten aus Video, Bild, Sprache und Text sind facettenreich und vielschichtig.[5, 6] Facettenreich bedeutet in diesem Zusammenhang, dass eine einzige Dateneinheit eine Vielzahl von Informationen enthalten kann. Welcher Informationsgehalt relevant ist, hängt von dem fachlichen Blickwinkel und der zugrundeliegenden Analysefrage ab. Der Informationsgehalt in Sprach- bzw. Sprechdaten besteht beispielsweise nicht nur aus dem gesprochenen Text, sondern auch aus der Tonhöhe oder anderen akustischen Eigenschaften. Das Merkmal der vielschichtigen Darstellung bedeutet, dass eine Dateneinheit mehr als einen Beitrag zur Information darstellen kann.

Die genannten Eigenschaften von unstrukturierten und nicht-standardisierten Daten führen zu einer Erweiterung des traditionellen Modells der empirischen Forschung (s. Abb. 1). Von seltenen Fällen abgesehen können unstrukturierte und nicht-standardisierte Daten nicht direkt in empirische Modelle einfließen. Sie müssen zunächst in eine strukturierte Form umgewandelt werden, die sich für eine weitere (kausale) Analyse eignet. Da die Daten sehr umfangreich sind und der spezifische Informationsgehalt nicht a priori festge-



1 Traditionelles und erweitertes Modell der empirischen Forschung

https://doi.org/10.3196/1864295020691280 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY-SA

legt ist, werden Algorithmen des maschinellen Lernens eingesetzt, um die Daten zu verdichten und in strukturierte Informationen umzuwandeln. Erst dann können Forschende die Daten für weitere empirische Analysen nutzen, um daraus wissenschaftliche Erkenntnisse zu generieren.

Strukturierte Daten und unstrukturierte Daten bilden gemeinsam das Datenuniversum für die empirische Forschung in den Wirtschafts- und Sozialwissenschaften. Das Zusammenspiel der beiden unterschiedlichen Datentypen spiegelt sich in der Struktur der NFDI in den beiden Konsortien KonsortSWD und BERD@NFDI, die komplementär und synergetisch zueinander konzipiert sind. So deckt KonsortSWD, das bereits seit 2020 im Rahmen der Nationalen Forschungsdateninfrastruktur gefördert wird und auf das vom RatSWD aufgebaute Netzwerk fußt, die Belange der strukturierten Daten in den Wirtschafts-, Sozial- und Bildungswissenschaften ab.[7] Die großen Herausforderungen bei der Arbeit mit strukturierten Daten sind die fragmentierte Datenlandschaft sowie rechtliche und ethische Fragen. Bei quantitativen Analysen von unstrukturierten, nicht-standardisierten Daten stellen sich andere Fragen: Die Art und Weise der Datenerhebung und -verwaltung ist aufgrund der Heterogenität der Quellen unterschiedlich, und die Datenformate sind kaum standardisiert. Werkzeuge und Methoden des maschinellen Lernens werden benötigt, um die Daten einzusammeln und (vor-) zu verarbeiten. Da die Verarbeitung unstrukturierter Daten häufig Rechen- und Speicherkapazität erfordert, die über die Ausstattung von Arbeitsplatzrechnern hinausgeht und zudem derzeit noch kein gemeinsamer Ablageort im Sinne eines Datenrepositoriums existiert, besteht ein Bedarf an Unterstützung von sowohl Speicher- als auch Rechenkapazitäten. Diese Lücke soll mit BERD@NFDI geschlossen werden.

Status Quo und community-spezifische Anforderungen

Um zum Projektstart von BERD@NFDI ein aktuelles Bild über die drängendsten Themen im Umgang mit unstrukturierten Daten zu erhalten, wurden im Juni 2021 Workshops organisiert, in denen Promovierende der Graduate School of Economics and Social Sciences (GESS) der Universität Mannheim zu diesem Themenkomplex befragt wurden. Unter den Teilnehmenden befanden sich Promovierende vom ersten Jahr bis zum vierten Jahr der Promotion. Alle Teilnehmenden hatten bereits gewisse Erfahrungen bei Projekten mit unstrukturierten Daten gemacht, wobei nicht alle Vorhaben auch durchgeführt wurden: Das ganze Spektrum von »in einem frühen Stadium verworfen« bis hin zu »erfolgreich abgeschlossen« war vertreten. Entlang des Forschungsdatenlebenszyklus stachen folgende Probleme besonders hervor, entweder, weil sie besonders häufig genannt wurden, oder aber, weil sie besonders grundlegender Natur auf dem Gebiet des Forschungsdatenmanagements sind:

Community-Aufbau & Support

Der Umgang mit unstrukturierten Daten erfordert umfangreiche IT- und Programmierkenntnisse, die in Studien- und Graduiertenprogrammen jedoch noch nicht in der Breite vermittelt werden. Passende Ansprechpartner*innen sind häufig nicht vorhanden, sodass beim Einstieg in die Thematik der unstrukturierten Daten hohen Hürden zu überwinden sind. Ein häufig genanntes Desiderat bei allen folgenden Punkten des Forschungsdatenlebenszyklus war es daher, eine Community aufzubauen, welche eine niedrigschwellige Diskussion über Fragen und Probleme mit anderen Forschenden und Expert*innen ermöglicht. Neben dem Austausch unter Forschenden wurde häufig auch Support von infrastruktureller Seite gewünscht, insbesondere bei der Datenerhebung, Rechtsthemen und der Nutzung der Cluster-Infrastruktur.

Datenerhebung

Unstrukturierte Daten entstehen in der Regel nicht originär für Forschungszwecke, und es gibt daher häufig auch keine standardisierten Zugangs- und Erhebungswege. Zudem kann der Datenumfang insbesondere bei Multimedia-Daten schnell sehr groß werden. Als wesentliche Problembereiche wurden somit die Ressourcenintensität der Datenerhebung (Zeit, Kosten, Rechner- und Speicherkapazität), Unsicherheit in rechtlichen Fragen, fehlende methodische Fachkenntnisse, Mehrfacharbeiten durch unkoordiniertes Vorgehen und ein fehlendes Verzeichnis existierender Datensätze (»Datenkatalog«) genannt. Gerade im Kontext der Betriebswirtschaftslehre ist die Forschung auch auf Daten von Unternehmen angewiesen, deren Beschaffung häufig aufgrund von rechtlichen Hürden ebenfalls sehr mühsam ist, sodass auch hierbei Support und Koordination als gewinnbringend angesehen werden.

Datenaufbereitung

Unstrukturierte Daten liegen selten in einer Form vor, die eine direkte Analyse mit Hinblick auf die Forschungsfrage erlaubt, sondern sie erfordern häufig umfangreiche Aufbereitungsschritte. Helfen würde den Forschenden im Bereich des Pre-Processing beispielsweise ein Erfahrungsaustausch und Unterstützung z. B. bei der Text- und Strukturerkennung, Computer Vision oder dem Mergen von Datensätzen, bzw. koordiniertes Vorgehen bei umfangreicheren Projekten.

Datenanalyse

Im Bereich der Analyse unstrukturierter Daten sind die zwei großen Themenkomplexe methodischer Support, z. B. durch Erfahrungsaustausch unter Peers, Unterstützung beim Auffinden und Bewerten von Analyseverfahren und deren Implementierung sowie ein niedrigschwelliger Zugang zu angemessenen Speicher- und Rechenkapazitäten.

https://doi.org/10.3196/1864295020691280 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY-SA

Datenarchivierung

Die Bedeutung eines langfristigen Speicherns, Erschließens und Erhaltens von Daten mit entsprechender Suchmöglichkeit wird von einer großen Mehrheit der Befragten bestätigt. Bibliotheken werden hierbei als kompetente Partner wahrgenommen, insbesondere im Bereich von Rechtsfragen und in praktischer Hinsicht. Gleichzeitig werden jedoch auch die Kosten einer sachgerechten langfristigen Aufbewahrung von Daten problematisiert und eine entsprechende Lösung im Zuge der NFDI nahegelegt.

Nachnutzen und Teilen von Daten

Der große Themenkomplex, der hier genannt wurde, betrifft rechtliche Fragen und Unsicherheiten, beispielsweise ob und wie Daten geteilt werden können, die kollaborativ in Projekten erhoben wurden, die von Unternehmen lizenziert wurden, unter den Datenschutz fallen oder die auf Grundlage der Text-und-Data-Mining-Regelungen des Urheberrechtsgesetzes analysiert wurden. Gleichzeitig wird auch das Fehlen von Anreizmechanismen zum Teilen von Daten beanstandet, selbst wenn dies eigentlich problemlos möglich wäre. Überlegungen zielten auf einen Top-down-Ansatz, bei dem z. B. der/die Betreuer*in der Doktorarbeit oder eine Promotionsordnung eine sachgemäße Dokumentation und Archivierung von Daten verpflichtend macht.

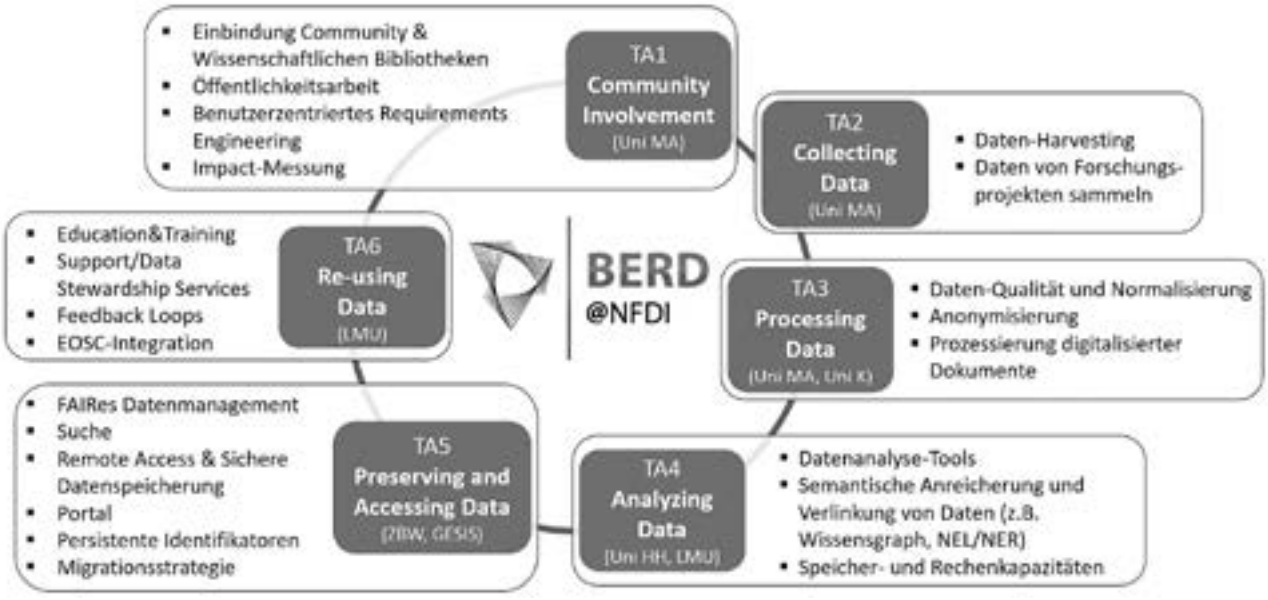
Zusammenfassend und ohne das Konzept beim Namen genannt zu haben, wird in der Breite der Teilnehmenden ein Umgang mit Daten befürwortet, der den FAIR-Kriterien entspricht. Integraler Bestandteil einer Forschungsdateninfrastruktur sind für die Teilnehmenden aber auch der Aufbau von Community- und Supportstrukturen sowie ein niedrigschwelliger Zugang zu ausreichender Hardware. Diesen Anliegen möchte sich BERD@NFDI in seinem Arbeitsprogramm widmen.

Arbeitsprogramm

Zentrales Anliegen von BERD@NFDI ist es, Forschende der Wirtschafts- und Sozialwissenschaften effektiv dabei zu unterstützen, unstrukturierte Daten zu erheben, zu speichern und zu analysieren. Der Aufbau, die Bereitstellung und die kontinuierliche Weiterentwicklung der entsprechenden Infrastrukturen, Services und Tools erstreckt sich über fünf inhaltliche Task Areas, die im Folgenden vorgestellt werden. Aufgrund der positiven Erfahrungen dem DFG-Projekt GeRDI,[8] welches eines der Vorläuferprojekte war, orientieren sich die Task Areas an dem Forschungsdatenmanagementzyklus, wie ihn das UK Data Archiv definiert hat.

Task Area 1: Community Involvement

Die Einbeziehung der Fachcommunity nimmt in BERD@NFDI eine zentrale Rolle ein. Ziel ist es, eine dauerhafte Verbindung zur Zielgruppe aufzubauen und die Nutzer*innen aktiv in das Vorhaben einzubinden, beginnend mit der Konzeption und dem Entwurf über die Umsetzung bis hin zur Bewertung des Projekts. Ein nutzerzentriertes Design und eine agile Entwicklungsmethodik sind wichtige Möglichkeiten, um die aktive Rolle der Fachcommunity im Projekt zu realisieren. So ist vorgesehen, jede Projektleistung mit Maßnahmen zur Verbreitung und Analyse der Nutzererfahrung zu begleiten. Dabei wird nicht nur die direkt am Projekt beteiligte Pilotgemeinschaft einbezogen, sondern auch



2 Task Areas und Maßnahmen im Arbeitsprogramm von BERD@NFDI

https://doi.org/10.3196/1864295020691280 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY-SA

weitere interessierte Communitys und potenzielle Interessengruppen.

Nachdem die Infrastrukturdienste von BERD@NFDI in Betrieb sind, wird ihre Nutzung und der Interaktion der Nutzer*innen mit den Diensten kontinuierlich überwacht, um zu ermitteln, welche Dienste und Funktionalitäten beliebt, ausbaufähig oder verbesserungswürdig sind. Schließlich sollen die Auswirkungen (Impact) des Projekts auf das Forschungsdatenmanagement innerhalb der Zielgruppe von BERD@NFDI gemessen werden.

Task Area 2: Collecting BERD

Der Forschungsdatenmanagementzyklus beginnt typischerweise mit dem Sammeln der Daten. Im Kontext von BERD@NFDI werden unstrukturierte Daten in der Regel nicht für Forschungszwecke erhoben, sondern entstehen nebenbei im Zuge der Datafication. Dies bringt mehrere Herausforderungen mit sich, die in den Community-Workshops zur Bedarfserhebung angesprochen wurden: Unstrukturierte Daten können unterschiedliche Formate (z. B. ASCII, PNG, MP3, MPEG-4) und Größenordnungen (von wenigen Megabyte bei Text bis hin zu mehreren GigaBytes bei Videos) umfassen. Entsprechend kommen Speicher- und Rechenkapazitäten von lokalen Ressourcen einer typischen Arbeitsplatzausstattung schnell an ihre Grenzen.

Auch in rechtlicher Hinsicht bestehen bei unstrukturierten Daten aus nicht-standardisierten Quellen häufig Unsicherheiten: Selbst wenn Text und Data Mining von Daten urheberrechtlich für Forschungszwecke möglich ist, so ist die Frage der Speicherung und Zugänglichmachung für Replikationen oder Anschlussforschung nach derzeitigem Stand unklar. Die Grenzen zwischen rechtlich zulässigem und unzulässigem Handeln sind dabei für Forschende häufig schwer erkennbar. So kann das Web Scraping von Daten aus Webseiten legal sein, die öffentliche Bereitstellung der so eingesammelten Daten kann wiederum illegal sein. Beispielsweise kann das Urheberrecht verletzt werden, wenn Fotos eingesammelt und ohne Zustimmung der Urheber*in an anderer Stelle öffentlich bereitgestellt werden. Und schließlich greift die Datenschutz-Grundverordnung, sobald personenbezogene Daten betroffen sind.

Vor diesem Hintergrund zielt diese Task Area darauf ab, eine robuste Infrastruktur bereitzustellen, über die unterschiedliche Methoden zum Einsammeln von Daten bereitgestellt werden. Darüber hinaus wird BERD@NFDI einen Prozess und Regeln zur Identifizierung, Bewertung und Priorisierung relevanter Quellen für strukturierte und unstrukturierte Datenquellen für die Forschung entwickeln. Schließlich sollen Forschende in die Lage versetzt werden, zu bewerten und zu entscheiden, welche Daten sie mit anderen Forschenden teilen dürfen und welche rechtlichen Restriktionen ggf. zu beachten sind.

Task Area 3: Processing BERD

Die in Task Area 2 eingesammelten Daten stammen häufig nicht von einer einzigen Stelle (Institution oder Community), sie unterliegen möglicherweise unterschiedlichen Datenmanagementpraktiken und können mit unterschiedlichen Datenerfassungsansätzen erstellt werden. Dies resultiert in unterschiedlichen Niveaus der Datenqualität, mit denen die Forschenden umgehen müssen. Das gleiche Problem stellt sich bei den Metadaten von Daten, die aus externen Quellen in die BERD@NFDI-Infrastruktur importiert werden. Wenn die Daten bereits in strukturierter Form vorhanden sind, können etablierte Prüf- und Normalisierungsverfahren genutzt werden, um die Qualität der (Meta-) Daten zu verbessern. Für unstrukturierte Daten können zwar neue Methoden der Klassifizierung, Normalisierung und Qualitätsbeurteilung angewandt werden, aber es gibt keinen allgemein akzeptierten Standard für die in BERD@NFDI abgedeckten Fachdisziplinen.

Um diese Situation zu verbessern, konzentriert sich diese Task Area zunächst auf die Entwicklung von Standards und Richtlinien für die Verarbeitung und Dokumentation von unstrukturierten Daten. Dies umfasst auch Maßnahmen zur Normalisierung von Metadaten. Um personenbezogene Daten ohne Offenlegung ihrer Identität entsprechend der Datenschutz-Grundverordnung zu verarbeiten, werden Anonymisierungsverfahren umgesetzt. Schließlich behandelt die Task Area die Verarbeitung digitalisierter Dokumente. So werden Werkzeuge angepasst und entwickelt, die helfen, unstrukturierten Text aus digitalen Bildern zu extrahieren. Dies wird besonders nützlich sein, um die Datenbasis für historische Ansätze in den Sozialwissenschaften und Wirtschaftswissenschaften zu vervollständigen.

Task Area 4: Analyzing BERD

Um substanzielle Forschungsfragen zu untersuchen, suchen die Forschenden nicht nur nach relevanten Daten, sondern auch nach relevanten Algorithmen zur Datenanalyse, insbesondere solchen aus dem Bereich der künstlichen Intelligenz und des maschinellen Lernens. BERD@NFDI wird eine breite Palette solcher Algorithmen bereitstellen – und ergänzend dazu eine Bewertung etwa der Genauigkeit (accuracy) und Präzision (precision), Lizenzinformationen, Testdatensätzen, auf denen die Verfahren trainiert wurden, oder Leistungsvergleichen, die auf einer transparenten Dokumentation der Verwendung von Algorithmen mit Datensätzen basieren.

Zur Beantwortung komplexer Forschungsfragen genügt es häufig nicht, einen einzelnen Datensatz zu analysieren. Vielmehr müssen vorhandene Daten mit anderen Daten aus weiteren Quellen angereichert und kombiniert werden. Dies kann sehr komplex sein, wenn gemeinsame Terminologien und Merkmale fehlen. Grundlegende Werkzeuge zur automatisierten Erken-

https://doi.org/10.3196/1864295020691280 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY-SA

nung normierter Entitäten in unstrukturierten Daten wurden bereits im Vorgängerprojekt BERD@BW entwickelt und als Open Source bereitgestellt.[9] BERD@NFDI wird ein erweitertes semantisches Netz bereitstellen, um komplexe Beziehungen zwischen Unternehmen und ihrem sozialen und historischen Umfeld zu erfassen.

Task Area 5: Preserving & Giving Access to BERD

Die Aufbewahrung und Dokumentation von sowie die Sicherstellung der Zugänglichkeit zu Daten ist eine zentrale Aufgabe von BERD@NFDI. Denn erst solche Funktionalitäten gewährleisten Reproduzierbarkeit, Wie-derverwendung, Analyse und Datenzitierung. Die Aufbewahrung digitaler Bestände erfordert eine regelmäßige Kontrolle der Daten und Metadaten, weshalb Strategien für die langfristige Gewährleistung der Datenintegrität, zusammen mit der Standardisierung von Metadaten und einem sicheren Speichermechanismus entscheidend sind. Aufbewahrungsrichtlinien spezifizieren die Anforderungen an die Datenspeicherung, die in der Regel Aspekte der Rechteverwaltung, Herkunfts- und Kontextinformationen und Zuständigkeiten umfassen. Außerdem hängt die Dauer der Aufbewahrung auch von verschiedenen Datenvalidierungs- und Handhabungsaktivitäten während des gesamten Lebenszyklus der Forschungsdaten ab. Dazu gehören auch Maßnahmen zur Gewährleistung der Datenintegrität, zur Pflege und Gewährleistung der Datenkonsistenz und -genauigkeit sowie eine sichere Datenspeicherung. Darüber hinaus ist eine Datenmigrationsstrategie, die die Schritte für die Verlagerung von Daten von einem System in ein anderes klärt, erforderlich. Sobald die Daten gesammelt und aufbewahrt wurden, ist die Bereitstellung des Zugangs gemäß den FAIR-Prinzipien ein weiteres wichtiges Merkmal der in BERD@NFDI zu entwickelnden Forschungsdateninfrastruktur.

Um diese Ziele zu erreichen, wird zunächst ein zentrales Eingangsportal bereitgestellt. Dahinter wird sich eine föderierte Dateninfrastruktur[10] befinden, deren Komplexität jedoch den Forschenden verborgen bleiben wird. Für die über BERD@NFDI bereitgestellten Metadaten ist ein gemeinsamer Metadatenstandard zu entwickeln.[11] Für diesen Zweck wird auf generischen Standards, wie DataCite[12] aufgebaut, um diese dann durch disziplinspezifische Standards wie dem der Data Documentation Initiative DDI[13] zu ergänzen. Darüber hinaus wird großer Wert auf die Einhaltung der FAIR-Prinzipien durch die Anwendung der FAIR Digital Objects Spezifikationen gelegt. Zudem werden unterschiedliche Services für die Suche bereitgestellt. Gängige Ansätze umfassen die Verwendung einer Suchmaschine zur Indizierung von Metadaten in einem zentralen Index und verschiedenen Suchstrategien, wie z. B. Entdeckung (discovery) oder Empfehlung (recommendation).

Task Area 6: Re-Using BERD

Um die Nachnutzung von Daten in den Wirtschafts- und Sozialwissenschaften zur gelebten Realität werden zu lassen, ist es wichtig, ein Umfeld schaffen, das den an der Datenproduktion Beteiligten den Wert der gemeinsamen Nutzung vor Augen führt. Die meisten Wissenschaftler*innen (und oft auch Behörden) haben nicht die Zeit oder das Geld, um Daten in einem für andere nutzbaren Format aufzubereiten oder Forschenden, die an der Analyse ihrer Daten interessiert sind, Unterstützung zu leisten. Daher zögern sie oft, die Daten überhaupt weiterzugeben. BERD@NFDI wird den zentralen Zielgruppen maßgeschneiderte Schulungen anbieten, um sie bei der Nachnutzung der relevanten Daten zu unterstützen. Darüber hinaus werden die organisatorischen, rechtlichen und technischen Voraussetzungen dafür geschaffen, dass die gewünschten Dienste ohne unnötige Barrieren genutzt werden können. Schließlich werden Werkzeuge für automatisierte Feedbackschleifen und Datenexportfunktionen geschaffen, um eine maximale Verbreitung und Auffindbarkeit der auf BERD@NFDI bereitgestellten Daten zu gewährleisten.

Feedbackschleifen begegnen dem Problem, dass Wissenschaftler*innen selten die Zeit aufbringen können bzw. wollen, umfassende Metadaten zu ihren verwendeten Daten zu erstellen. Eine effizientere und nachhaltigere Strategie ist die direkte Extraktion dieser Informationen aus Veröffentlichungen. Zum Beispiel beschreiben die Daten- und Methodenabschnitte in einer Publikation ausführlich die Besonderheiten des Datensatzes, welche Variablentransformationen vorgenommen wurden und warum und wie die Daten kombiniert und gewichtet wurden. Diskussionsabschnitte heben Unzulänglichkeiten der vorhandenen Daten hervor. Wenn diese Daten systematisch gesammelt werden, können sie von enormem Wert sein: Andere Nutzer*innen der gleichen oder verwandter Daten erfahren, was bereits erprobt und getestet wurde, und Datenproduzenten und -anbieter erhalten Hinweise, um ihre Datenqualität zu verbessern.

Die Rolle der beteiligten Bibliotheken

Innerhalb des BERD@NFDI-Konsortiums kommt den beiden beteiligten Bibliotheken eine zentrale Rolle zu. Sie bringen nicht nur ihre Expertise in Bezug auf Forschungsdatenmanagement, Entwicklung und dauerhaften Betrieb von Infrastrukturen, dem Management von Metadaten, Standardisierung, Education usw. ein, sondern sie sind vielmehr die zentralen Akteure im übergreifenden operativen und technischen Projektmanagement. Dies entspricht ihrem Selbstverständnis von Bibliotheken als verlässlichen und zugleich innovativen Partnern der Wissenschaft auf Augenhöhe.

Dienstleistungen zu Forschungsdaten stellen einen wesentlichen Bestandteil des Serviceportfolios der Uni-

https://doi.org/10.3196/1864295020691280 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY-SA

versitätsbibliothek Mannheim dar. So betreibt sie seit 2013 das institutionelle Repositorium MADATA[14] als lokale Forschungsdateninfrastruktur, und über da|ra besteht seitdem eine Kooperation zur Registrierung von DOIs mit der ZBW und GESIS. Mit dem Aktienführer-Datenarchiv[15] bietet die Universitätsbibliothek Mannheim eine überregionale Forschungsdateninfrastruktur an, über die Daten zu tausenden an deutschen Börsen notierten Aktiengesellschaften zwischen 1956 und 2019 zur Verfügung gestellt werden.[16] Das Projekt legte auch den Grundstein für die Erweiterung der Digitalisierungsexpertise[17] in Richtung optische Zeichenerkennung (OCR) sowie maschinelle Text- und Layoutanalyse. Zwischenzeitlich ist die UB in mehreren Projekten zur Weiterentwicklung von OCR-Verfahren und -Tools involviert, z. B. im Rahmen von OCR-D und OCR-BW.[18] Diese Expertise der Prozessierung unstrukturierter digitalisierter Dokumente bringt die UB Mannheim in BERD@NFDI ein.

Zur Weiterentwicklung in Richtung einer echten semantischen Aufbereitung und Repräsentation von wirtschaftswissenschaftlichen Forschungsdaten engagiert sich die Universitätsbibliothek Mannheim seit gut zwei Jahren in dem Aufbau eines Wissensgraphen mit Daten rund um deutsche Firmen.[19,20] Diese Wissensbasis wird die Universitätsbibliothek Mannheim in BERD@NFDI auf mehrere Millionen Einträge zu deutschen Firmen weiter ausbauen und als Tool zum Recherchieren, Datenverknüpfen und -anreichern sowie zum Download zur Verfügung stellen. Die Forschungsdatenaktivitäten werden im universitätsweiten Forschungsdatenzentrum gebündelt, das in der Universitätsbibliothek angesiedelt ist.

Darüber hinaus zeichnet sich die Universitätsbibliothek Mannheim durch eine starke Nutzer- und Serviceorientierung aus – mit jährlich über 2,3 Millionen Besuchen, rund 4,5 Millionen Downloads auf Journals und E-Books und 2 Millionen Datenbankabfragen.[21] Insbesondere die Erfahrung in der persönlichen Beratung sowie der Vermittlung von Informationskompetenz über physische und virtuelle Kanäle ist eine wesentliche Stärke der Universitätsbibliothek Mannheim,[22] die auch in BERD@NFDI innerhalb der Maßnahme »Training and Education« zum Tragen kommt.[23] So wurden an der Universitätsbibliothek Mannheim bereits Data-Literacy-Kurse sowohl auf prä- als auch postgradualer Ebene konzipiert und umgesetzt und damit das Engagement im Forschungsdatenmanagement erfolgreich mit den Aktivitäten der Informationskompetenz verschränkt. Die hohe Serviceorientierung prädestiniert die Universitätsbibliothek Mannheim für die Aufgabe der zentralen Projektkoordination. Sie verantwortet sämtliche übergreifenden organisatorischen, verwaltungs- und finanztechnischen Prozesse des BERD@NFDI-Konsortiums.

Die ZBW entwickelt und betreibt seit vielen Jahren eigene Infrastrukturen, wie beispielsweise das Rechercheportal EconBiz mit ca. 11 Millionen Titelnachweisen und mit jährlich über 1 Million eindeutigen Besucher*innen sowie das Open Access Repository EconStor mit über 9 Millionen Downloads im Jahr 2020. Diese Kompetenz ist vor dem Hintergrund, dass in BERD@NFDI eine dauerhafte, robuste und auf die Bedarfe der Zielgruppe zugeschnittene Forschungsdateninfrastruktur aufgebaut wird, unerlässlich. Daher wird auch der Chief Technology Officer (CTO) für BERD@NFDI von der ZBW gestellt.

Das Management von Metadaten ist eine weitere bibliothekarische Kompetenz, die für BERD@NFDI von zentraler Bedeutung ist. Die ZBW hat in zahlreichen vergangenen Projekten moderne Ansätze zum Metadatenmanagement entwickelt, insbesondere in föderierten Umgebungen, wie sie im Bereich Forschungsdaten häufig anzutreffen sind (vgl. Endnote 8). Zu nennen sind hier Methoden zur Entwicklung von gemeinsamen Metadatenstandards für Ressourcen aus unterschiedlichen Quellen sowie deren modulare und disziplinären Erweiterung (vgl. Endnote 9). Von besonderer Bedeutung sind in diesem Zusammenhang auch die Expertise im Kontext der FAIR-Prinzipien.[24] Die ZBW betreibt das deutsche Büro der internationalen GO FAIR Initiative, die bereits im vierten Jahr Fragestellungen zur Umsetzung der FAIR-Prinzipien in unterschiedlichsten Kontexten begleitet. Damit ist sichergestellt, dass BERD@NFDI die FAIR-Prinzipien auf dem jeweils neuesten Stand der Entwicklungen berücksichtigen wird. Schließlich können Bibliotheken als Gedächtnisorganisationen in der Regel auf eine lange Historie in der Pflege von nationalen, europäischen und internationalen Netzwerken zurückblicken. Die ZBW hat ihre ursprünglichen, primär bibliothekarischen Netzwerke um Netzwerke im Bereich Forschungsdatenmanagement erweitert. Dadurch ist sichergestellt, dass BERD@NFDI z. B. neueste Entwicklungen aus dem Rat für Informationsinfrastrukturen in Deutschland (RfII), der im Aufbau befindlichen europäischen Forschungsdateninfrastruktur »European Open Science Cloud« (EOSC)[25] und den bedeutendsten internationalen Forschungsdateninitiativen, wie der »Research Data Alliance« (RDA),[26] des »World Data System« (WDS),[27] des »Committee on Data of the International Science Council« CODATA[28] und GO FAIR[29] unmittelbar aufgreifen kann.

Vor dem Hintergrund dieser Darstellungen wird zum einen deutlich, auf Basis welcher Expertise Bibliotheken zu bedeutenden Partnern in NFDI-Konsortien werden. Zum anderen zeigt dies aber auch auf, welche vielfältigen Entwicklungsperspektiven Bibliotheken haben, um sich inhaltlich weiterzuentwickeln.

https://doi.org/10.3196/1864295020691280 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY-SA

Anmerkungen

1 DRUMMER, Alan. Extracting insights from complex, unstructured big data. In: *Journey to AI Blog*. 2020 [Zugriff am: 29. November 2021]. Verfügbar unter: https://www.ibm.com/blogs/journey-to-ai/2020/11/managing-unstructured-data/

2 BERD steht für Business, Economic and Related Data.

3 Partner des Science Data Centers BERD@BW sind das Mannheim Center for Data Science (MCDS), die Universitätsbibliothek und die Universitäts-IT der Universität Mannheim sowie der Forschungsbereich »Innovationsökonomik und Unternehmensdynamik« und das Forschungsdatenzentrum des Zentrums für Europäische Wirtschaftsforschung ZEW. Gefördert wird das Science Data Center vom Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg im Rahmen der Landesdigitalisierungsstrategie seit 2019 und bis voraussichtlich Ende 2022.

4 LYCETT, Mark. ›Datafication': making sense of (big) data in a complex world. In: *European Journal of Information Systems*. 2013, 22 (4), S. 381–386. DOI https://doi.org/10.1057/ejis.2013.10.

5 BALDUCCI, Bitty und Detelina MARINOVA. Unstructured data in marketing. In: *Journal of the Academy of Marketing Science*. 2018, 46 (4), S. 557–590. DOI https://doi.org/10.1007/s11747-018-0581-x.

6 GANDOMI, Amir und Murtaza HAIDER. Beyond the hype: Big data concepts, methods, and analytics. In: *International Journal of Information Management*. 2015, 35 (2), S. 137–144. DOI https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

7 Vgl. den Beitrag dazu im vorliegenden Heft, S. 48–58.

8 LATIF, Atif, Fidan LIMANI und Klaus TOCHTERMANN. A Generic Research Data Infrastructure for Long Tail Research Data Management. In: *Data Science Journal*. Ubiquity Press, 2019, 18 (1), S. 17. DOI https://doi.org/10.5334/dsj-2019-017.

9 spaCyOpenTapioca – A spaCy wrapper of OpenTapioca for named entity linking on Wikidata [Zugriff am: 7. Dezember 2021]. Verfügbar unter: https://github.com/UB-Mannheim/spacyopentapioca

10 LATIF, Atif, Fidan LIMANI und Klaus TOCHTERMANN. On the Complexities of Federating Research Data Infrastructures. In: *Data Intelligence*. 2021, 3 (1), S. 79–87. DOI https://doi.org/10.1162/dint_a_00080.

11 LIMANI, Fidan, Atif LATIF, Timo BORST, u. a. Metadata Challenges for Long Tail Research Data Infrastructures. In: *Bibliothek Forschung und Praxis*. 2019, 43 (1), S. 68–74. DOI https://doi.org/10.1515/bfp-2019-2014.

12 DataCite Schema [Zugriff am: 7. Dezember 2021]. Verfügbar unter: https://schema.datacite.org/

13 Data Documentation Initiative (DDI) [Zugriff am: 7. Dezember 2021]. Verfügbar unter: https://ddialliance.org/

14 MADATA – Mannheim research data repository [Zugriff am: 7. Dezember 2021]. Verfügbar unter: https://madata.bib.uni-mannheim.de/

15 GEHRLEIN, S., KAMLAH, J., PINTSCH, M., SCHUMM, I. und WEIL, S. (2020). Vom Papier zur Datenanalyse. »Neue« historische Forschungsdaten für die Wirtschaftswissenschaften. In: *E-Science-Tage 2019: Data to Knowledge* (S. 140–152). heiBOOKS: Heidelberg.

16 Für die Nutzung der Datenbank, welche im Rahmen eines DFG-Projekts zwischen 2013-2019 entstanden ist, haben sich zwischenzeitlich 175 Institutionen und über 600 Personen freischalten lassen.

17 HÄNGER, Christian, Irene SCHUMM und Stefan WEIL. Alte Drucke in neuem Gewand: ein Beispiel für den erfolgreichen Einsatz der freien Digitalisierungsplattform Goobi an der UB Mannheim. In: *BIT online: Zeitschrift für Bibliothek, Information und Technologie mit aktueller Internet-Präsenz*, 2015, 18 (3), S. 231–239.

18 WEIL, Stefan und Jan KAMLAH. OCR-BW – Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen: Texterkennung von historischen Drucken mit OCR-D und Tesseract. Online, 2020 [Zugriff am: 7. Dezember 2021]. Verfügbar unter: https://madoc.bib.uni-mannheim.de/57424; WEIL, Stefan und Jan KAMLAH. Forschungsdaten aus Digitalisaten. In: Vincent HEUVELINE (Hrsg.). *E-Science-Tage 2019: Data to Knowledge*. Bd. 598. Heidelberg: heiBOOKS, 2019, S. 189. Verfügbar unter: https://doi.org/10.11588/heibooks.598.c8429

19 Implementiert wird der BERD-Wissensgraph in der offenen Software Wikibase, mit der auch der größte existierende Wissensgraph Wikidata betrieben wird. Zudem engagiert sich die Universitätsbibliothek Mannheim in der Weiterentwicklung der Wikibase, bspw. als Mitglied der Wikibase-Stakeholder-Group, und durch die (Weiter-)Entwicklung von Tools zur effektiven Arbeit mit der Wikibase; siehe auch Ontology, Knowledge Graph and Reconciliation Services | BERD@NFDI. 2021 [Zugriff am: 6. Dezember 2021]. Verfügbar unter: https://www.berd-nfdi.de/servicestools/knowledge-graph/

20 SHIGAPOV, Renat. RaiseWikibase: Towards fast data import into Wikibase. Online, 2021 [Zugriff am: 7. Dezember 2021]. Verfügbar unter: https://madoc.bib.uni-mannheim.de/60059; SHIGAPOV, Renat, Philipp ZUMSTEIN, Jan KAMLAH, u. a. bbw: Matching CSV to Wikidata via Meta-lookup. In: Ernesto JIMÉNEZ-RUIZ, Oktie HASSANZADEH, Vasilis EFTHYMIOU, u. a. (Hrsg.). *CEUR Workshop Proceedings*. Bd. 2775. Aachen: RWTH, 2020, S. 17–26. Verfügbar unter: https://madoc.bib.uni-mannheim.de/57386

21 Die Angabe der Downloads und Datenbankabfragen bezieht sich auf das Jahr 2020. Die Besuche-Statistik bezieht sich auf das Jahr 2019 vor Corona.

22 In einem durchschnittlichen Jahr beantwortet die Universitätsbibliothek gut 70.000 Anfragen an den Infotheken, und bietet über 400 Kursstunden sowie 220 Veranstaltungen mit 4.500 Teilnehmer*innen an.

23 S. Abschnitt »Task Area 6: Re-Using BERD«.

24 DREFS, Ines; LINNE, Monika; TOCHTERMANN, Klaus: FAIRe Forschung. Wie Wissenschaftliche Bibliotheken den Herausforderungen von Open Science begegnen. In: *BuB – Forum Bibliothek und Information* 2018, 70 (11), S. 636–639; http://hdl.handle.net/11108/387

25 EOSC Association [Zugriff am: 7. Dezember 2021]. Verfügbar unter: https://eosc.eu/

26 RDA | Research Data Sharing without barriers [Zugriff am: 7. Dezember 2021]. Verfügbar unter: https://www.rd-alliance.org/

27 World Data System: Trusted Data Services for Global Science [Zugriff am: 7. Dezember 2021]. Verfügbar unter: https://www.worlddatasystem.org/

28 CODATA – The Committee on Data for Science and Technology [Zugriff am: 7. Dezember 2021]. Verfügbar unter: https://codata.org/

29 GO FAIR initiative: Make your data & services FAIR [Zugriff am: 30. November 2021]. Verfügbar unter: https://www.go-fair.org/

https://doi.org/10.3196/1864295020691280 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY-SA

Verfasser*innen

Dr. Sabine Gehrlein, Direktorin,
Universitätsbibliothek Mannheim,
Schloss Schneckenhof West, 68131 Mannheim,
Telefon +49 621 181-2941,
sabine.gehrlein@bib.uni-mannheim.de
Foto: Sebastian Weindel

Dr. Annette Klein, Stellvertr. Bibliotheksdirektorin,
Universitätsbibliothek Mannheim,
Schloss Schneckenhof West, 68131 Mannheim,
Telefon +49 621 181-2975,
annette.klein@bib.uni-mannheim.de
Foto: Sebastian Weindel

Dr. Irene Schumm, Leitung Forschungsdatenzentrum | UB-Projektleitung BERD-BW,
Universitätsbibliothek Mannheim,
Schloss Schneckenhof West, 68131 Mannheim,
Telefon +49 621 181-2754,
irene.schumm@bib.uni-mannheim.de
Foto: Sebastian Weindel

Prof. Dr. Klaus Tochtermann, Direktor, ZBW –
Leibniz-Informationszentrum Wirtschaft,
Düsternbrooker Weg 120, 24105 Kiel,
Telefon +49 431 8814-333,
k.tochtermann@zbw.eu
Foto: ZBW, Sven Wied

https://doi.org/10.3196/1864295020691280 https://www.inlibra.com/de/page/agb - Open Access - (cc) BY-SA