Auszug aus dem Buch:

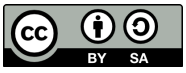
Uwe Jensen
Sebastian Netscher
Katrin Weller (Hrsg.)

Forschungsdatenmanagement sozialwissenschaftlicher Umfragedaten

Grundlagen und praktische Lösungen für den Umgang mit quantitativen Forschungsdaten

Verlag Barbara Budrich
Opladen • Berlin • Toronto 2019

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über
http://dnb.d-nb.de abrufbar.



Dieses Buch steht im Open-Access-Bereich der Verlagsseite zum kostenlosen Download bereit (https://doi.org/10.3224/84742233).
Eine kostenpflichtige Druckversion (Print on Demand) kann über den Verlag bezogen werden. Die Seitenzahlen in der Druck- und Onlineversion sind identisch.

ISBN 978-3-8474-2233-4 (Paperback)
eISBN 978-3-8474-1260-1 (eBook)
DOI 10.3224/84742233

Umschlaggestaltung: Bettina Lehfeldt, Kleinmachnow – www.lehfeldtgraphic.de
Lektorat: Nadine Jenke, Potsdam
Satz: Anja Borkam, Jena – kontakt@lektorat-borkam.de
Titelbildnachweis: Foto: Florian Losch
Druck: paper & tinta, Warschau
Printed in Europe

12. Räumliche Verknüpfung georeferenzierter Umfragedaten mit Geodaten: Chancen, Herausforderungen und praktische Empfehlungen

Stefan Müller

Georeferenzierte Daten sind Daten, die mit direkten Raumbezügen, d.h. Geokoordinaten angereichert wurden (Meyer/Bruderer Enzler 2013: 323). Anwendungen für diese Daten finden sich vor allem in wissenschaftlichen Fachdisziplinen wie der Ökologie, z.B. bei der Untersuchung natürlicher Habitate von Vögeln oder der Bodenbeschaffenheit von Waldgebieten (Plant 2012: 9ff.). In der sozialwissenschaftlichen Umfrageforschung ist seit einigen Jahren ebenfalls eine verstärkte Nachfrage (Schweers et al. 2016; RatSWD – Rat für Sozial- und Wirtschaftsdaten 2012) sowie ein zunehmender Einsatz (Bluemke et al. 2017; Hillmert/Hartung/Weßling 2017) georeferenzierter Umfragedaten zu beobachten. Die Hoffnung von Forschenden ist, dass sich durch die kleinräumige Verortung von Befragten sowie die räumliche Verknüpfung dieser Orte mit interessanten Nachbarschaftsmerkmalen die Kontexte sozialen Handelns besser erfassen und verstehen lassen (Stimson 2014: 18). In der Tat finden sich umfangreiche Arbeiten in den verschiedensten Teilbereichen der sozialwissenschaftlichen Forschung, etwa in der Analyse von politischen Verhalten und Einstellungen (Förster 2018; Klinger/Müller/Schaeffer 2017), sozialen Bedingungen von Gesundheit (Saib et al. 2014) oder sozialräumlichen Einflüssen auf Bildungsübergänge (Weßling 2016).

Die Nutzung von georeferenzierten Umfragedaten hat indessen weitreichende Implikationen im Bereich des Forschungsdatenmanagements. Es müssen technische, organisatorische, datenschutzrechtliche und dokumentarische Fragestellungen geklärt werden. Denn zum einen handelt es sich bei der Nutzung von georeferenzierten Umfragedaten um ein interdisziplinäres Unterfangen (Dietz 2002: 540), das Forschenden und Forschungsprojekten Kenntnisse der Daten und entsprechender Software zu ihrer Nutzung abverlangt (Meyer/Bruderer Enzler 2013: 319). Zum anderen sind georeferenzierte Daten vor allem sehr sensible Daten (Skinner 2012: 8), die im Sinne des Datenschutzes besonders geschützt werden müssen. Schließlich sollten die Prozesse der Datenerhebung und räumlichen Verknüpfung angemessen dokumentiert werden – ein Unterfangen, das bei interdisziplinären Projekten oftmals erschwert ist (Edwards et al. 2011: 669f).

Diesen Herausforderungen im Forschungsdatenmanagement ist das vorliegende Kapitel gewidmet, in welchem sie systematisiert und anhand praktischer Lösungsmöglichkeiten diskutiert werden. In Abschnitt 12.1 werden zunächst grundlegende Begriffe und Prozesse geklärt. Die drei darauffolgenden Abschnitte widmen sich den technischen sowie organisatorischen (12.2), datenschutzrechtlichen (12.3) und dokumentarischen (12.4) Herausforderungen. Zunächst werden dazu jeweils die einzelnen Herausforderungen vorgestellt und anschließend Lösungsmöglichkeiten diskutiert. In Abschnitt 12.5 wird schließlich die Weitergabe von georeferenzierten Umfragedaten zur Sekundäranalyse skizziert, bevor Abschnitt 12.6 das vorliegende Kapitel zusammenfasst.

https://doi.org/10.3224/84742233.13

12.1 Begriffe und wichtigste Prozesse

Wie eingangs erwähnt haben georeferenzierte Umfragedaten eine interdisziplinäre Komponente. Diese folgt daraus, dass etwa die Ökonomie, Soziologie oder Geographie über unterschiedliche Werkzeuge und Terminologien zur Beantwortung ihrer Forschungsfragen verfügen (Dietz 2002: 540). Forschende und Forschungsprojekte aus den Sozialwissenschaften sind folglich bei der Arbeit mit Daten aus anderen Disziplinen häufig mit Begriffen und Prozessen konfrontiert, die ihnen fremd sind und einer Klärung bedürfen. Im Folgenden werden daher die wichtigsten Begriffe und Prozesse vorgestellt.

12.1.1 Georeferenzierung, Geokodierung und Geodaten

Als georeferenzierte Umfragedaten werden oft Umfragedaten bezeichnet, denen standardisierte Raumbezüge zugeordnet wurden. Zu diesen standardisierten Raumbezügen zählen Namen und Bezeichner (ID) für räumliche Einheiten wie Kreise, Gemeinden oder Postleitzahlengebiete (Hillmert et al. 2017: 270f.). Diese Zuordnung hat u.a. große Vorteile für die Analyse der Daten. So können etwa Abhängigkeiten zwischen Befragten, verursacht durch Klumpungen von Personen im Raum, kontrolliert werden oder Kontextdaten aus anderen Quellen, z.B. über gemeinsame Namen und Bezeichner, den Umfragedaten hinzugefügt werden.

In den Geowissenschaften wie z.B. der Ökologie ist der Begriff der Georeferenzierung hingegen enger gefasst: Demnach bezeichnet Georeferenzierung die Zuordnung von Geokoordinaten zu Daten (RatSWD 2012: 11). Im Gegensatz zu Namen und Bezeichnern sind Geokoordinaten Koordinatenpunkte, die über ein entsprechendes Koordinatensystem, das über die Erdoberfläche gespannt wurde, die Lage eines Punktes auf jener Erdoberfläche verorten. Die Zuordnung von Geokoordinaten zu Daten hat den Vorteil, dass Beobachtungen im Raum zueinander in Beziehung gesetzt werden können und explizite Analysen basierend auf diesem Raumbezug möglich werden. Dazu gehören z.B. die Berechnung von geographischen Distanzen zwischen Punkten oder die Errechnung von Flächenanteilen umliegender Flächen einzelner Punkte (Meyer/Bruderer Enzler 2013: 327ff.).

Für eine Georeferenzierung müssen also Geokoordinaten vorliegen. Oft ist es dafür notwendig, die Daten mit indirektem Raumbezug (beispielsweise Adressen von Befragten) in Geokoordinaten zu übersetzen. Diesen Vorgang der Umwandlung nennt man Geokodierung. Hierzu können automatisierte Dienste genutzt werden, die auf Datenbanken zugreifen, welche Adressinformationen sowie zugehörigen Geokoordinaten beinhalten und diese wechselseitig konvertieren (Zandbergen 2014: 2). Allerdings wird dabei die Geokodierung selten von einzelnen Forschungsprojekten lokal, d.h. am eigenen Computer, vorgenommen, da das Betreiben eines solchen Dienstes technisch sehr aufwändig ist. Daher ist oft die Inanspruchnahme von Drittanbietern von Geokodierungsdiensten wie zum Beispiel Google, Bing oder dem Bundesamt für Kartographie und Geodäsie (BKG) erforderlich.

Im Kontext der Georeferenzierung sind schließlich noch Geodaten (im Allgemeinen) zu erwähnen. Geodaten sind Daten, deren Informationen durch Geokoordinaten im Raum dargestellt und analysiert werden können. Diese Informationen und deren zugrunde liegenden Beobachtungseinheiten nehmen je nach Geodatensatz unterschiedliche Ausdehnungen im Raum in Form von Geometrien ein, wie z.B. ein Punkt oder ein Polygon. Geodaten und deren enthaltene Geometrien sind somit bereits georeferenziert. Zwar können auch georeferenzierte Umfragedaten aus den Sozialwissenschaften als Geodaten bezeichnet werden, es ist jedoch sinnvoll, georeferenzierte Umfragedaten von Geodaten explizit zu unterscheiden. Aus daten-

schutzrechtlichen Gründen müssen Geokoordinaten und Umfragedaten stets getrennt gespeichert werden. Daher liegen Umfragedaten zumindest auf Adressebene nie als georeferenzierte Daten im Sinne von eigenständigen Geodatensätzen vor, wie im Abschnitt 12.3 ausführlicher erörtert.

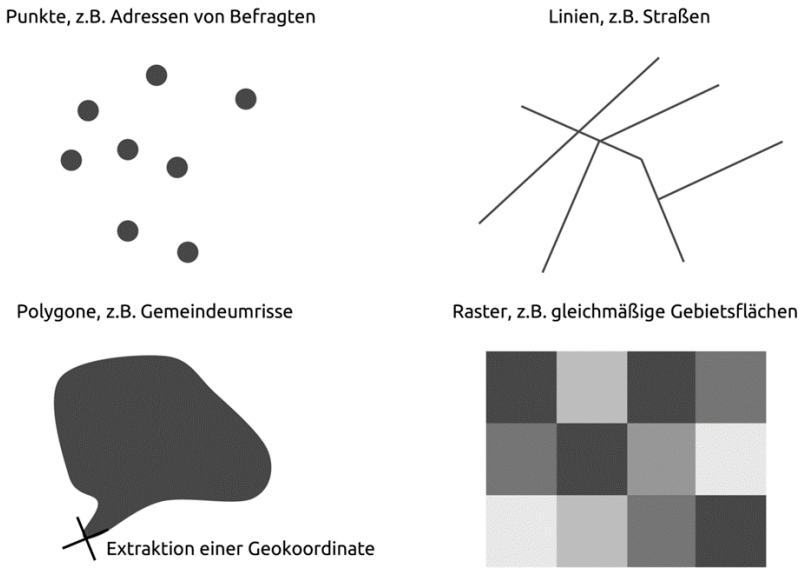
12.1.2 Räumliche Verknüpfung

Eine räumliche Verknüpfung ist die Verbindung zweier zuvor getrennt vorliegender georeferenzierter Datensätze bzw. Geodaten. Dazu wird zunächst eine Zielquelle (z.B. der georeferenzierte Datensatz A) ausgewählt, der räumliche Eigenschaften (Geometrien) oder Attribute (Fachinhalte) einer anderen Quelle (z.B. der georeferenzierte Datensatz bzw. Geodatensatz B) hinzugefügt werden. Wie in Abbildung 12.1 dargestellt, sind Geometrien eine geometrische Form der räumlichen Einheiten, wie etwa:

- Punkte für Hausadressen von Befragten,
- Linien für Straßen,
- Polygone für Stadteilumrisse oder
- gleichmäßige Rasterflächen für grenzübergreifende Merkmale.

Fachinhalte sind inhaltliche Eigenschaften, die diese Geometrien als Attribute enthalten können: Anzahl der Personen, die an einer Adresse wohnen (= Punkte), Hauptverkehrsstraßen oder Nebenstraßen (= Linien), Anzahl der Arbeitslosen in einem Stadtteil (= Polygone) oder die Luftverschmutzung über Stadtgrenzen hinweg (= Rasterflächen). Durch die Georeferenzierung ist es möglich, einen willkürlichen Punkt innerhalb einer Geometrie zu wählen und eine Geokoordinate für diesen Punkt zu extrahieren, wie es in Abbildung 12.1 anhand eines Punktes der Polygon-Geometrie dargestellt ist. Dies ist die Grundlage für räumliche Verknüpfungen.

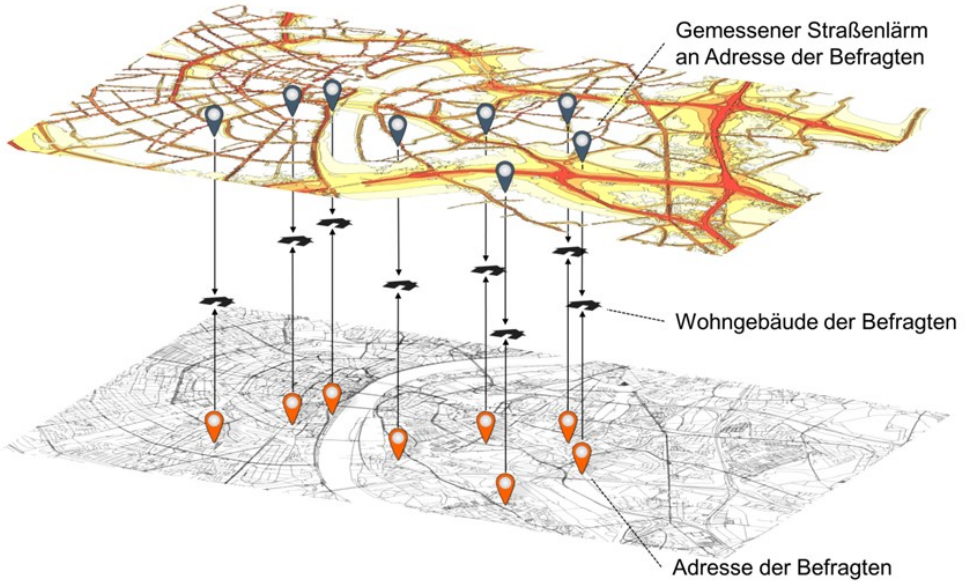
Abbildung 12.1: Mögliche Geometrien verschiedener Geodaten



Quelle: Eigene Darstellung

Denn durch die Projektion in einem gemeinsamen Koordinatenraum können in der Folge räumliche Verknüpfungen realisiert werden. Abbildung 12.2 zeigt diesen Vorgang anhand des Beispiels von Umgebungslärmdaten und Geokoordinaten von fiktionalen Teilnehmenden an einer Umfrage. Die Abbildung zeigt zwei Karten: eine Karte auf der unteren Ebene, die Straßen in Form von Linien abbildet, und eine Karte auf der oberen Ebene, die mit den Straßen assoziierten Verkehrslärm in Form von Polygonen darstellt. Dadurch, dass beide Karten in einem gemeinsamen Koordinatenraum projiziert werden, kann für jeden beliebigen Punkt einer gewählten Karte die Information der jeweilig anderen Karte extrahiert werden. Somit lässt sich beispielsweise analysieren, auf welchen Straßen ein Dezibelwert von mehr als 50 gemessen wurde.

Abbildung 12.2: Räumliche Verknüpfung von geokodierten Adressdaten und Straßenlärmdaten



Quellen: Eigene Darstellung unter Nutzung von EIONET Data Repository (CDR) für die Straßenlärmdaten und OpenStreetMap für die Straßenkarte (OSM).

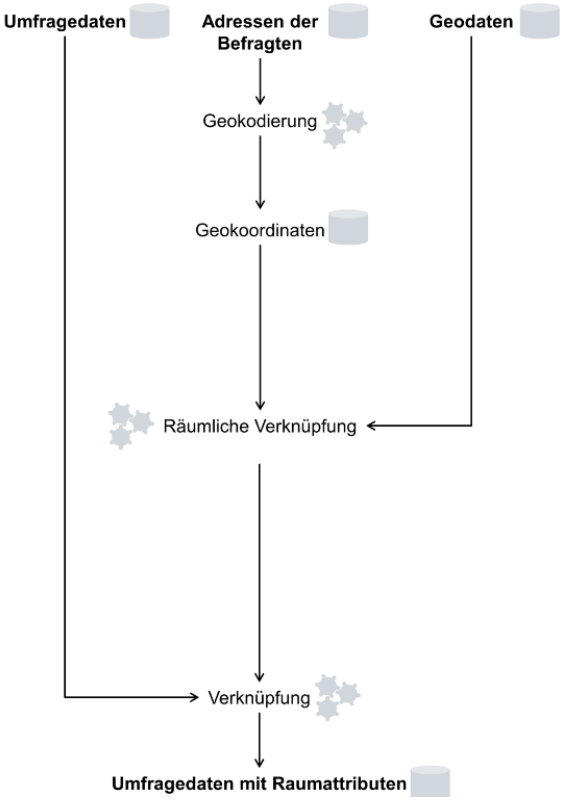
Orangene Marker stellen die geokodierten Adressen dar, welche gemeinsam mit den Straßenlärmdaten in einen gemeinsamen Raum projiziert werden. Graue Marker repräsentieren die entsprechenden gemessenen Werte des Straßenlärms an der geokodierten Adresse.

So können nicht nur Geometrien verbunden, sondern auch die den Geometrien zugeordneten Attribute, wie z.B. Verkehrslärm, verknüpft werden. Ferner können diese Attribute durch die Projektion in einem Koordinatenraum verschiedentlich bearbeitet werden, was durch eine Reihe standardisierter Verfahren möglich wird (Strobl 2017: 472). Ein gängiges Beispiel ist die Berechnung von Distanzen, wie etwa die Berechnung für Koordinatenpunkte, für die keine Straßenlärmmessung vorliegt, oder die Distanz zur nächsten Messung eines Dezibelwerts einer bestimmten Höhe. Analog lässt sich auch die mittlere gemessene Lärmbelastung in einem Umkreis von z.B. 100 Metern abbilden. Mit herkömmlichen Methoden über gemeinsame Namen oder Identifikatoren, wie etwa Gemeindeschlüssel oder Postleitzahlen, lassen sich derartige Verknüpfungen nicht vornehmen – einerseits aus Mangel einer kleinräumigen Verortung im Raum, anderseits aufgrund der fehlenden Projektion der Beobachtungen als Geometrien in einem gemeinsamen Koordinatenraum.

12.1.3 Exemplarischer Verlauf der räumlichen Verknüpfung

Georeferenzierte Umfragedaten haben die Eigenschaft, dass es sich dabei um Individualdaten handelt, die in Verbindung mit Adressinformationen personenbezogen sind. Nach der aktuellen Datenschutzgesetzgebung (BDSG) in Deutschland, die auf der Datenschutzgrundverordnung der EU (DSGVO) und dem Gesetz zur Anpassung des Datenschutzrechts an die Verordnung (DSAnpUG-EU) basiert, müssen gemäß § 27 Abs. 3 (DSAnpUG-EU) Merkmale gesondert gespeichert werden, „mit denen Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren Person zugeordnet werden können". Diese Merkmale sind u.a. Adressen, zu denen somit auch auf Adressen beruhende Geokoordinaten gehören. Aus diesem Grund hat sich ein Prozess der räumlichen Verknüpfung etabliert, welcher eine strikte Trennung der verwendeten Datensätze vorsieht. Abbildung 12.3 stellt diesen Prozess exemplarisch dar.

Abbildung 12.3: Ablauf der räumlichen Verknüpfung georeferenzierter Umfragedaten mit Geodaten



Quelle: Eigene Darstellung

Zunächst existieren drei voneinander getrennt gespeicherte Datenquellen: die Umfragedaten, die dazugehörigen Adressinformationen und die Geodaten. Ziel einer räumlichen Verknüpfung ist es, Informationen aus den Geodaten über die Adressinformationen der Befragten den Umfragedaten hinzuzufügen. Dazu werden die Adressdaten geokodiert und mit den Geodaten räumlich verknüpft. Erst dann werden diese neu hinzugewonnenen Informationen mit den eigentlichen Umfragedaten verknüpft, sodass zuletzt ein Umfragedatensatz mit soge-

nannten Raumattributen vorliegt. Die dafür notwendigen Zwischenschritte werden im Detail in den folgenden Abschnitten erörtert.

12.2 Technische und organisatorische Aspekte der räumlichen Verknüpfung

Die Arbeit mit georeferenzierten Umfragedaten und ihrer räumlichen Verknüpfung mit Geodaten bedarf einiger technischer sowie organisatorischer Vorbereitungen, nicht zuletzt aufgrund der bereits diskutierten Interdisziplinarität. Forschungsprojekte müssen sich daher sehr früh mit technischen und organisatorischen Herausforderungen auseinandersetzen, da sich je nach konkretem Forschungsvorhaben große Implikationen seitens des Ressourcenmanagements ergeben können.

12.2.1 Herausforderungen: Technik, Software und Ressourcen

Zu den zwei wichtigsten technischen und organisatorischen Herausforderungen einer räumlichen Verknüpfung gehören im Wesentlichen der angemessene Umgang mit zwei verschiedenen Datentypen – (georeferenzierte) Umfragedaten und Geodaten – sowie der Einsatz entsprechender Software, mit der sich diese beiden Daten verbinden lassen.

Doch wie unterscheiden sich zunächst die beiden Datentypen? Umfragedaten werden zumeist in einer flachen rechteckigen Datenstruktur in Form von einfachen Tabellen, der sogenannten Datenmatrix, erfasst. Einer Konvention folgend stellen die Zeilen dieser Tabellen die Beobachtungen oder Fälle dar und die Spalten die sogenannten Variablen oder Attribute, die sich aus den kodierten Antworten einer Befragung ergeben. Geodaten lassen sich zwar in bestimmten Dateiformaten ebenfalls in einer rechteckigen Form darstellen, z.B. als Comma Separated Text Files (CSV), die Weiterverarbeitung und Analyse führt jedoch weg von der flachen Datenstruktur hin zu einer mehrdimensionalen Struktur. So kann grundsätzlich zwischen Informationen hinsichtlich der Geometrien (z.B. Punkte oder Polygone) und den Fachinhalten (Höhe der Luftverschmutzung an einem Messpunkt, Anteil von Kindertagesstätten in einem Stadtteil) unterschieden werden. Beobachtungen aus einem Geodatensatz sind somit einerseits durch ihre Lage im Raum sowie durch weitere Informationen, die mit ihnen attribuiert sind, beschrieben.

Um mit diesen Daten zu arbeiten, wird spezielle Software eingesetzt: *Geographische Informationssysteme* (GIS). Die Handhabung von GIS muss aufgrund ihrer Komplexität von Forschenden erlernt werden (Meyer/Bruderer Enzler 2013: 319). Es handelt sich dabei um Software zur Bearbeitung, Analyse, aber auch graphischen Darstellung raumbezogener Daten (Bluemke et al. 2017). Daneben müssen je nach Projektkontext auch gewisse Hardwareanforderungen bewältigt werden. Geodaten können u.U. sehr groß sein. Je nach Umfang und Dateiformat sind Dateigrößen im Gigabyte-Bereich keine Seltenheit. Zuletzt werden für die eigentliche Bearbeitung dieser Daten daher seitens der Computerausstattung angemessene Prozessor-Taktungen und eine ausreichende Verfügbarkeit von Arbeitsspeicher benötigt.

12.2.2 Antworten und Lösungen: Personal-, Weiterbildungs- und Ressourcenpolitik

Die Antworten für Forschungsprojekte auf diese technischen und organisatorischen Herausforderungen sind zunächst sehr einfach: Es bedarf einer angemessenen Weiterbildungs-, Personal- und Ressourcenplanung, die optimalerweise bereits vor Projektbeginn vorgenommen werden sollte. Allerdings fallen die insbesondere hinsichtlich der Ressourcenplanung bestehenden Details je nach Projektkontext unterschiedlich kompliziert aus. Aus diesem Grund kann hier keine Blaupause darüber erstellt werden, welche Maßnahmen und Kosten wie und in welchem Maße z.B. bei der Erstellung von Forschungsdatenmanagementplänen anfallen. Es gibt jedoch Abwägungsüberlegungen, die im Folgenden vorgestellt werden und entsprechend des konkreten Projektvorhabens spezifiziert werden müssen.

Relativ nahe liegt zunächst die eigene Weiterbildung oder gezielte Anwerbung von Mitarbeitenden mit entsprechenden Fähigkeiten im Umgang mit Geoinformationssystemen. Dies ist oft nötig, da GIS in den Geistes- und Sozialwissenschaften noch verhältnismäßig wenig Verwendung finden (Meyer/Bruderer Enzler 2013: 319) – auch wenn neuere Trends darauf hindeuten, dass sich dieser Umstand im Wandel befindet (Bluemke et al. 2017: 309).

Mitarbeitende müssen schließlich über die nötige Ausstattung verfügen können. Das heißt Forschungsprojekte müssen sich, je nachdem welche konkreten Forschungsvorhaben anstehen, mit der Lizenzierung von Software etwa von Geoinformationssystemen auseinandersetzen. Denn Geoinformationssysteme sind stark auf dem kommerziellen Markt vertreten, wie z.B. das Produkt *ArcGIS* der Firma ESRI (2015). Mittlerweile gibt es zwar gut nutzbare freie Software, wie z.B. *QGIS* (QGIS Development Team 2018), und für Nutzende in den Sozialwissenschaften ist wahrscheinlich insbesondere die Nutzung der freien Software *R* interessant (R Core Team 2017). Aber obwohl sich räumliche Verknüpfungen somit schon mit relativ geringem Ressourcenaufwand bewältigen lassen (Müller/Schweers/Siegers 2017), sind gerade größere Projekte auf die Unterstützung durch Softwarefirmen bei Skalierungs- und Performanzproblemen angewiesen. Hier ist eine kritische Evaluierung der jeweiligen Systemanforderungen für die einzusetzende Software hinsichtlich des konkreten Forschungsvorhabens notwendig.

Neben der Erweiterung der eigenen (institutionellen) Fähigkeiten oder Ressourcen können räumliche Datenverknüpfungen alternativ auch institutionell auslagert werden. Diese Auslagerung fasst im besten Fall die Anforderungen an die erforderliche fachliche Expertise und die benötigte technische Ausstattung zusammen. Tatsächlich existiert ein relativ umfangreicher Markt an Anbietern. Oft haben diese einen Schwerpunkt auf Marktforschung, wie z.B. die Firmen microm und Infas 360, die zum einen die notwendige Geokodierung und zum anderen die Verknüpfung mit vielfältigen kleinräumigen Geodaten anbieten. Dazu gehören soziodemographische Merkmale eines Wohnviertels, wie Alter, Familienstände und Ausländeranteile, oder auch sozioökonomische Indikatoren, wie z.B. detaillierte Angaben zu Kaufkraftindizes.

Analog steigt aber auch die Anzahl nicht kommerzieller Anbieter (Schweers et al. 2016: 107ff.), sodass mittlerweile viele öffentliche Daten als frei verfügbare Geodaten angeboten werden. Beispielsweise existieren mit der Geodateninfrastruktur Deutschland (GDI-DE) oder dem GOVDATA Portal im Internet frei verfügbare, harmonisierte Geodatenangebote, die eine Vielzahl von Daten aus den unterschiedlichsten Fachdisziplinen anbieten. Auf europäischer Ebene soll hier vor allem die Initiative *Infrastructure for Spatial Information in Europe* (INSPIRE) erwähnt werden, die von einer starken Zunahme an europaweiten harmonisierten Geodaten in den kommenden Jahren ausgeht. Obwohl diese Angebote zum jetzigen Zeitpunkt nicht derart umfassend sind wie jene kommerzieller Anbieter, stellen sie bereits heute eine gute Quelle für in der Forschung nutzbare Geodaten dar (Förster 2018; Klinger/Müller/Schaeffer 2017). Und während bei kommerziellen Anbietern oft das Black-Box-

Prinzip gilt, d.h. die Datengenese ist ein Betriebsgeheimnis, existiert gerade bei öffentlichen Daten eine Pflicht zur Transparenz. Allerdings, und hier liegt der Unterschied zu kommerziellen Anbietern, finden sich im öffentlichen Sektor wenige bis gar keine Anbieter, die ein vollständiges räumliches Verknüpfungsprojekt inklusive der Geokodierung von Adressen begleiten können. Kurzum: In der Regel können externe Anbieter die Geokodierung, Aufbereitung der Geodaten und die Verknüpfung mit den Geokoordinaten übernehmen. Die Pflege und Aufbereitung der Daten während und vor allem nach Abschluss des Projekts obliegt jedoch weiterhin den Forschungsprojekten. Das betrifft u.a. auch die Dokumentation der Daten, worauf weiter unten noch näher eingegangen wird.

Zusammenfassend bedeuten die technischen und organisatorischen Herausforderungen einer räumlichen Verknüpfung vor allem eines: einen erhöhten Ressourcenaufwand. Räumliche Verknüpfungsprojekte können sehr einfach durchgeführt werden, z.B. wenn die hinzugefügten Geoinformationen aus frei verfügbaren und harmonisierten Quellen stammen. Sie können jedoch auch erschwert werden, wenn diese beispielsweise nicht harmonisiert oder sogar fehlerhaft vorliegen (Schweers et al. 2016: 109f.). Forschungsprojekte sollten daher sehr genau evaluieren, welche Daten, Verknüpfungen und Analysen sie für ihr Forschungsvorhaben benötigen.

Wie die Forschungsprojekte diesem erhöhten Ressourcenaufwand begegnen, kann dabei sehr unterschiedlich ausfallen: entweder durch eine gezielte Personal-, Weiterbildungs- sowie technische Ausstattungsstrategie im Rahmen der Forschungsförderung oder durch Kooperation mit Dateninfrastrukturen bzw. die Nutzung von Dienstleistungen öffentlicher oder kommerzieller Anbieter von Geodaten. Die Entscheidung für das eine oder andere bzw. einen Mix kann indessen nicht universell beantwortet werden.

12.3 Datenschutz und Re-Identifikationsrisiko

In räumlichen Datenverknüpfungsprojekten kommt dem Thema Datenschutz eine besondere Bedeutung zu, da hier ggf. in allen Phasen der Speicherung, Geokodierung, Verknüpfung, Distribution und Analyse entsprechender Daten mit personenbezogenen bzw. sensiblen Informationen gearbeitet wird. Natürlich darf nicht unterschätzt werden, dass auch einfache Umfragedaten sensible Informationen beinhalten können. Gerade aber die Herstellung eines expliziten Raumbezugs von Umfragedaten verstärkt diese Problematik, da die räumliche Verortung die Zahl der für eine Re-Identifikation infrage kommenden Personen stark eingrenzt. Beispielsweise ist die Identität einer Anwältin mit sieben Kindern in einem bestimmten Stadtteil einer bekannten Stadt wesentlich einfacher zu bestimmen, als eine Anwältin mit sieben Kindern, von der man nur weiß, in welchem Bundesland sie lebt.

12.3.1 Herausforderungen: räumliche Verknüpfung und zusätzliche Informationen

Die datenschutzrechtlichen Herausforderungen bei der Arbeit mit georeferenzierten Daten können abermals auf zwei Ebenen angetroffen werden: einerseits bezüglich des Verknüpfungsprozesses sowie der notwendigen technischen Vorarbeit, wie eben erörtert; andererseits durch die Verknüpfung und das Hinzufügen zusätzlicher Attribute zu den Umfragedaten.

Bezogen auf den ersten Herausforderungskomplex steht die angestrebte Verknüpfung georeferenzierter Umfragedaten mit Geodaten zunächst vor dem Problem, dass nach den gültigen Datenschutzregeln Adressinformationen, und damit auch Geokoordinaten, nicht gemein-

sam mit Umfrageattributen gespeichert werden dürfen. Aus diesem Grund werden in typischen Umfrageprojekten Adress- und Umfragedaten in separaten, bestenfalls lokal getrennten Dateien vorgehalten.[1] Oftmals übernimmt gar das beauftragte Erhebungsinstitut die Speicherung der Adressen. Je nach Größe des Projekts haben dann jeweils verschiedene Mitarbeitende Zugriff auf die separaten Daten.

Gleichzeitig muss spätestens für die räumliche Verknüpfung dieser Daten eine Korrespondenz hergestellt werden. So ist es durchaus möglich, die Geokoordinaten der Befragten einer Umfrage mit Informationen aus weiteren Geodaten (z.B. Dezibelwerte aus Verkehrslärmdaten) räumlich zu verknüpfen. Doch wenn diese Geokoordinaten der Befragten nun nicht mit den eigentlichen Umfrageinformationen gemeinsam gespeichert werden dürfen oder sollten, wie sollen die Dezibelwerte für einen gemeinsamen Analysedatensatz räumlich verknüpft werden? Diese Frage kann nur durch organisationale Abläufe des eigentlichen Verknüpfungsprozesses beantwortet werden.

Denn auch schon vor der eigentlichen räumlichen Verknüpfung müssen Fragen beantwortet werden, die sich aus der Geokodierung ergeben. Geokodierungsservices verarbeiten Daten nicht lokal, sondern bieten die Möglichkeit, über das Internet individuelle Geokoordinaten für individuelle Adressinformationen abzufragen (Zandbergen 2014: 2). Nutzende dieser Dienste laden z.B. eine CSV-Datei mit Adressinformationen zu dem jeweiligen Dienst hoch und können diese, nachdem der Prozess beendet wurde, angereichert mit Geokoordinaten wieder herunterladen.

Die technische Implementierung der Speicherung der Anfragen unterscheidet sich indessen je nach Dienst gravierend. Kommerzielle Anbieter speichern u.U. Anfragen an den Geokodierungsdienst, d.h. die Adressen der Befragten, was zu schwerwiegenden Problemen mit datenschutzrechtlichen Vorgaben führt. Es ist z.B. davon auszugehen, dass ein Unternehmen wie Google, dessen Geschäftsmodell auf Daten basiert, jegliche Anfragen an ihren Geokodierungsservice nachhält. Zu bedenken ist auch, dass datenschutzrechtlich zertifizierte Geokodierungsdienste wie jener vom BKG in der Regel nur Bundeseinrichtungen, also Bundesbehörden oder vom Bund finanzierte Forschungseinrichtungen, zur Verfügung stehen. Inwieweit diese geschützten Dienste wie z.B. der BGK GeoCoder auf vertraglicher Grundlage auch für (öffentlich finanzierte) Forschungszwecke genutzt werden können, sollte dort erfragt bzw. beantragt werden.

Zwar ist es durchaus möglich, einen eigenen Geokodierungsserver z.B. über die Datenbank der freien Kartenanwendung OpenStreetMap zu betreiben. Dieser sogenannte OpenStreetMap-Nominatim-Server ist jedoch sehr aufwendig zu pflegen. Zudem lassen sich mit OpenStreetMap nicht immer adressgenaue Geokodierungen vornehmen, da für bestimmte Adressen keine Hausnummern vorliegen.

Auch wenn Geokodierungsservices gefunden werden, die Anfragen nur flüchtig verarbeiten und nicht längerfristig speichern, z.B. über vertragliche Vereinbarungen, birgt der Weg über einen Webdienst Gefahren. Während Adressen für sich genommen nicht zwingend

1 In dem seit dem 25. Mai 2018 gültigen Gesetz zur Anpassung des Datenschutzrechts an die Verordnung steht allerdings auch der Passus, dass personenbezogene Daten (d.h. auch Adressen bzw. Geokoordinaten), „mit den Einzelangaben nur [sic!] zusammengeführt werden [dürfen], soweit der Forschungs- oder Statistikzweck dies erfordert“ (DSAnpUG-EU § 27 Abs. 3). Eine räumliche Verknüpfung könnte u.U. als ein solcher Forschungs- oder Statistikzweck bezeichnet werden. Allerdings war dieser Passus in der alten Fassung des BDSG bereits im gleichen Wortlaut enthalten. Dennoch hat sich in Forschungsprojekten der konservative Ansatz bewährt, diese Merkmale schlichtweg zu trennen. Im Folgenden wird zudem ein Verfahren vorgestellt, das eine räumliche Verknüpfung von georeferenzierten Umfragedaten mit Geodaten auch unter einer solchen strengen Regelung möglich macht.

personenbezogene Daten darstellen,[2] können sie sehr wohl in Verbindung mit weiteren Informationen Rückschlüsse auf Befragte, d.h. natürliche Personen, zulassen. Dazu gehören Metainformationen wie z.B. der Projekttitel einer Befragung im Dateinamen einer Adressliste, welche für den Geokodierungsservice genutzt wird, oder die korrespondierende Anzahl der Zeilen in dieser Datei mit der Anzahl der Ausschöpfungsquote jener Befragung. Auch kann der über IP-Adressen ermittelte Ort, von wo aus die Anfrage an den Dienst gestellt wurde, auf etwaig mit dem Ort verbundene Forschungsprojekte verweisen.

Die zweite Herausforderung bezüglich des Datenschutzes betrifft das Hinzufügen zusätzlicher Attribute zu den Umfragedaten. Gerade die Verortung im Raum, insbesondere im kleinräumigen Maßstab, schafft ein erweitertes Re-Identifikationsrisiko von Befragten. Auch die Attribute des Raums selbst – z.B. die Ausländeranteile, die Anteile von Arbeitslosen oder die Luftverschmutzung in einer bestimmten räumlichen Einheit, etwa einer Nachbarschaft – können einzigartige Beobachtungen in einem Datensatz erzeugen. In der Kombination mit anderen Attributen ist man so u.U. schnell in der Lage, Rückschlüsse auf einzelne Personen zu ziehen. Kapitel 4 geht im Rahmen der Anonymisierung näher auf dieses Problem der Re-Identifikation natürlicher Personen in Forschungsdaten ein.

Prinzipiell liegt die besondere Gefahr der Verknüpfung mit Geodaten darin begründet, dass Geodaten ebenso wie die Umfragedaten in maschinenprozessierbarer Form vorliegen. Zusatzinformationen zur Nachbarschaft der Befragten ergeben sich daher nicht aus vermeintlichem Insiderwissen, sondern durch Daten, die in aufbereiteter, durchsuchbarer und systematisch auswertbarer Form vorliegen. Dieser Umstand wird verstärkt, wenn es sich dabei um Geodaten handelt, die frei zugänglich sind und deren Informationen zu einem Umfragedatensatz hinzugefügt wurden.

12.3.2 Antworten und Lösungen: technische und organisationale Verfahren

Den beiden Herausforderungen bezüglich des Datenschutzes georeferenzierter Umfragedaten – getrennte Speicherung und Re-Identifikationsrisiko – kann im Wesentlichen durch einen vorab definierten Ablauf der räumlichen Verknüpfung begegnet werden. Zentrale Elemente betreffen sowohl die Auswahl geeigneter Ablageorte und Speicherverfahren als auch die organisationale Struktur der Verknüpfung selbst. Indessen muss die konkrete datenschutzkonforme Implementierung an das jeweilige Forschungsvorhaben angepasst und entsprechend umgesetzt werden. Abbildung 12.4 zeigt exemplarisch den vollständigen Ablauf der Georeferenzierung und räumlichen Verknüpfung, wie er u.a. für Forschungsprojekte angewendet werden kann, in welchen die Umfragedaten selbst erhoben werden und somit Zugriff auf die Adressdaten der Befragten besteht.

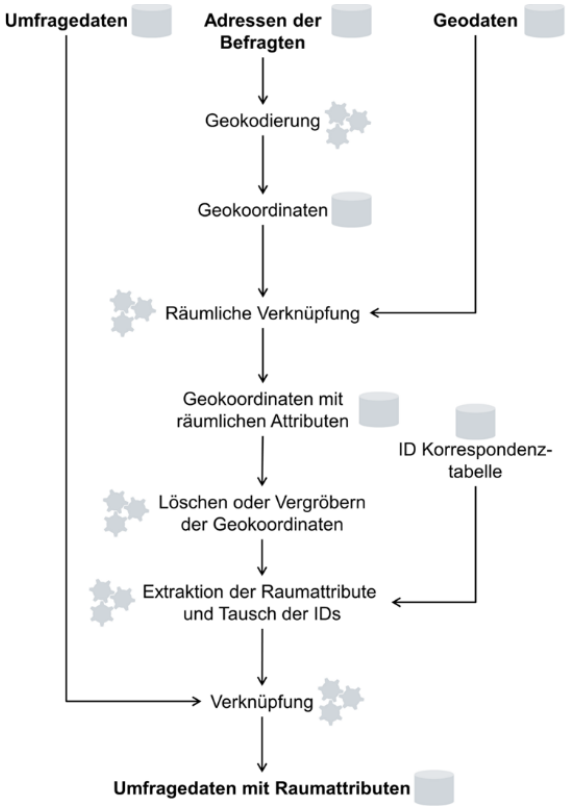
Zunächst wird im Ablauf der Georeferenzierung und der Datenverknüpfung sichergestellt, dass Umfragedaten (Datensatz A) und Adressen bzw. Geokoordinaten der Befragten (Datensatz B) zu keinem Zeitpunkt gemeinsam gespeichert werden. Dies wird erreicht, indem getrennte physische Ablageorte gewählt werden. In der Folge liegen die Umfragedaten auf Computer/Server A und die Adressinformationen auf Computer/Server B getrennt vor. Zusätzliche Sicherheit bietet der Einsatz eines dritten Ablageorts, Computer/Server C. Auf diesem wird eine Korrespondenztabelle der Identifikatoren, z.B. laufende Nummern für die Befragten in den Datensätzen A und B, gespeichert, welche sich jedoch zwischen diesen beiden Datensätzen unterscheiden. Somit ermöglicht die Korrespondenztabelle die

2 Adressen sind öffentlich verfügbare Informationen, die zunächst Gebäude referenzierbar machen. Allerdings kann beispielsweise die Adresse von einem alleinstehenden Gebäude, in welchem nur eine Person wohnt, sehr wohl direkte Informationen über die Identität einer einzelnen Person liefern.

Verknüpfung der Umfrage- mit den Adressdaten. Der Zugriff auf diese Korrespondenztabelle muss daher auch besonders streng gestaltet werden.

Bis zu dem Punkt, an welchem Geoinformationen und Umfragedaten in einen gemeinsamen Datensatz überführt werden, können alle Daten relativ problemlos aufbereitet und gemanagt werden. Nachdem die Adressdaten geokodiert wurden, können diese mittels der räumlichen Verknüpfung mit den Geodaten verknüpft werden. Im Verlauf müssen aber einige je nach Problemstellung unterschiedliche Vorkehrungen und Entscheidungen getroffen werden.

Abbildung 12.4: Vollständiger Ablauf der räumlichen Verknüpfung von georeferenzierten Umfragedaten mit Geodaten unter besonderer Berücksichtigung des Datenschutzes



Quelle: Eigene Darstellung

Erstens müssen, um die hinzugewonnenen Geoinformationen mit den Umfragedaten zu verknüpfen, die im verknüpften Datensatz noch enthaltenen Geokoordinaten gelöscht werden. Dadurch wird die direkte Re-Identifikation von befragten Personen über die Geokoordinaten vermieden. Zweitens können ggf. Rauminformationen der Geokoordinaten vergröbert werden, etwa durch eine Aggregation der Koordinaten auf hierarchisch höher gelagerte Raumgrenzen, wie z.B. Stadtteile. Dieser Schritt würde garantieren, dass trotz der Löschung der kleinräumigen Rauminformationen die statistische Kontrolle räumlicher Abhängigkeiten in späteren Analysen der Daten möglich bleibt. Ebenso lassen sich somit ex post zusätzliche Geoinformationen auf diesem höheren Aggregationsniveau hinzufügen.

Es existieren verschiedene Verfahren wie Rauminformationen, zu denen Geokoordinaten zählen, vergröbert werden können. Hierbei sollte grob zwischen zwei Verfahren unterschieden werden: erstens Verfahren, welche die ursprüngliche Geokoordinate verfremden (Kroll/Schnell 2016; Zandbergen 2014) und somit keinen Rückschluss mehr auf die Ursprungsgeokoordinate zulassen sollten (Kounadi/Leitner 2014); sowie zweitens Verfahren, die sich mit dem Re-Identifikationsrisiko in bereits vergröberten Raumeinheiten, wie z.B. Gemeinden oder Postleitzahlengebiete, auseinandersetzen (Blatt 2012; El Emam 2006). Dieser Bereich ist im Detail sehr komplex und je nach hinzugefügter Rauminformation kann die Entscheidung, welche Daten wie vergröbert oder verfremdet werden, sehr unterschiedlich ausfallen. In diesem Kapitel wird daher auf eine weitergehende Diskussion verzichtet.

Der Datenschutz von Befragten spielt zu jedem Zeitpunkt eines jeden Forschungsvorhabens mit georeferenzierten Umfragedaten eine große Rolle. Angefangen von der Geokodierung über die eigentliche Verknüpfung bis hin zur Weitergabe und somit Sekundärnutzung muss stets bedacht werden, dass mit sehr sensiblen Informationen gearbeitet wird, die besonderer Vorkehrungen bedürfen. Mit dem nötigen Problembewusstsein und einer entsprechenden Anpassung der Arbeitsabläufe und -prozesse lassen sich die mit georeferenzierten Umfragedaten verbundenen Risiken jedoch wirkungsvoll minimieren.

Es muss allerdings bedacht werden, dass die hier vorgestellten Lösungen der datenschutzbezogenen Herausforderungen sich vor allem auf die interne Anwendung innerhalb von individuellen Forschungsprojekten beziehen. Die Kontrolle über die Prozesse ist wesentlich erschwert, wenn Geodatenfachexpertise ausgelagert wird und Geokodierungen, räumliche Verknüpfungen und Analysen durch externe Dienstleister durchführt werden. Hier gilt es Versicherungen über Datenschutzvorkehrungen bei den jeweiligen Drittanbietern einzuholen. Letztlich muss der Weg über Drittanbieter jedoch nicht zwingend ein Problem sein, sondern kann datenschutzbezogene Vorkehrungen sogar vereinfachen. Ein entsprechendes Beispiel ist das Sozio-oekonomischen Panel (SOEP), dessen datenerhebendes Institut als Treuhändler der Adressdaten und ihrer Geokodierung auftritt. Die Primärforschenden des SOEP haben hingegen keinen Zugriff auf die Adressdaten der Befragten (Goebel/Wagner/Wurm 2010: 5).

12.4 Dokumentation georeferenzierter Umfragedaten

Kapitel 6 dieses Buches diskutiert die Bedeutung einer detaillierten Planung der Prozesse der Dokumentation von Forschungsdaten. Die Dokumentation georeferenzierter Umfragedaten unterscheidet sich nicht wesentlich von jener klassischer sozialwissenschaftlicher Umfrage- bzw. Forschungsdaten. Allerdings müssen einige Fallstricke beachtet werden, die sich daraus ergeben, dass Metadaten aus verschiedenen Fachdisziplinen zusammengebracht werden, die sich u.U. nicht mit dem etablierten sozialwissenschaftlichen DDI-Metadatenstandard abbilden lassen.

12.4.1 Herausforderungen: Metadaten aus verschiedenen Quellen

In vielen wissenschaftlichen Fachdisziplinen werden zur Dokumentation der Forschungsdaten bewährte Metadatenstandards genutzt. So ist in den Sozialwissenschaften der Metadatenstandard der Data Documentation Initiative (DDI) stark verbreitet (Vardigan 2013; Zenk-Möltgen 2012), wie in Kapitel 9.1 beschrieben. In den Geowissenschaften und öffentlichen

Geodateninfrastrukturen (GDI) wird hingegen der für Geodatendaten spezifische ISO-19115-Metadatenstandard genutzt (AdV/KLA 2015). Bei der Dokumentation georeferenzierter Daten müssen daher sowohl die räumliche Verknüpfung als auch die hinzugewonnenen Informationen inhaltlich beschrieben werden. Das ist dann erschwert, wenn sich die verfügbaren Metadatenfelder in einem fachspezifischen Metadatenstandard nicht an den Erfordernissen aus der jeweilig anderen Fachdisziplin orientieren.

Das Beispiel des DDI-Metadatenstandards ist dafür exemplarisch. So ermöglicht der DDI-Metadatenstandards etwa die Integration von Feldern aus dem Geodaten-Metadatenstandard ISO 19115 (Vardigan/Heus/Thomas 2008: 109). Entsprechende Felder in DDI sind explizit vorgesehen und erlauben beispielsweise die Beschreibung von geographischen Strukturen der Polygon-, Punkt- oder Rastergeometrien von Geodaten. Ebenso ist die Anwendung kontrollierter inhaltlicher Vokabulare aus ISO 19115 in DDI möglich.

Allerdings lassen sich nicht alle Informationen auf die jeweiligen Entitätsebenen in DDI übertragen. So können die bereits erwähnten geographischen Strukturen gegenwärtig nicht auf der sogenannten Variablenebene der Umfragedaten beschrieben werden. Das geht nur über die sogenannte Studienebene, welche Informationen über die Daten als solche bereithält und alle darin enthaltenen Variablen einbezieht. Somit können in Umfragedaten etwaig vorkommende geographische Strukturen nur für alle in den Daten enthaltenen Variablen gemeinsam beschrieben werden. Das wäre insofern kein Problem, solange räumliche Datenverknüpfungen nur zwischen den Umfragedaten und einer einzigen Geodatenquelle vorgenommen werden.

In der Praxis ist das jedoch kein realistisches Szenario. Ein Beispiel aus dem GESIS Datenarchiv für Sozialwissenschaften macht dies deutlich. Im Jahre 2015 wurden im Zuge der Georeferenzierung der Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften 2014 (GESIS 2015) deren Daten zunächst mit kleinräumigen Verkehrslärmdaten verknüpft. Ferner wurden durch die gleiche Methode kleinräumige Geoinformationen aus dem deutschen Zensus 2011 hinzugefügt. Auch konnte durch die Verortung von Befragten in einem Geokoordinatenraum und mit Hilfe von Daten des BKG und Gebietsänderungslisten des Statistischen Bundesamtes (destatis) Harmonisierungen der Gemeinden bis ins Jahr 1994 vorgenommen werden (Klinger 2018: 50ff.). Dabei stammten alle drei Geodaten aus drei verschiedenen Quellen und hatten verschiedene geographische Strukturen als Grundlage:

- ungleichmäßige Polygone der Verkehrslärmdaten,
- gleichmäßige 1 km² Rasterzellen der Zensusdaten sowie
- großflächige Polygone der Gemeindeumrisse.

Folglich konnten nicht alle drei Quellen zu Dokumentationszwecken in einer gemeinsamen inhaltlichen Kategorie zusammengefasst werden.

Wie Forschende sich derartigen Herausforderungen stellen können, widmet sich der nächste Unterabschnitt, in welchem Möglichkeiten der Dokumentationen für georeferenzierte Umfragedaten erörtert werden. Diese Möglichkeiten reichen von einfachen Workarounds bis hin zu komplexeren Verfahren, die jedoch u.U. zu Inkompatibilitäten mit bestehenden Standards führen.

12.4.2 Antworten und Lösungen: Workarounds und Brute-Force-Ansätze

Obwohl der DDI-Metadatenstandard in den Sozialwissenschaften weit verbreitet ist, ist die Möglichkeit, den ISO-19115-Metadatenstandard in DDI zu integrieren, gegenwärtig nur auf die Studienebene begrenzt. Um dem Problem fehlender Dokumentationsmöglichkeiten in DDI auf Variablenebene zu begegnen, bestehen verschiedene Möglichkeiten, die ursprüng-

liche Implementierung von ISO 19115 in DDI in unterschiedlichem Maße zu modifizieren. So besteht erstens die Option von *Workarounds*, die zwar der ursprünglichen Spezifikation des Standards widersprechen, aber keine Inkompatibilität mit dem Standard als solchen darstellen. Sogenannte *Brute-Force-Ansätze* stellen eine zweite Option dar, die zu eindeutigen Inkompatibilitäten führen, indem bewusst die Implementierung der ISO-19115-Metadaten auf Variablenebene – entgegen des DDC-L Konzepts – erzwungen wird. Beide Ansätze haben Vor- und Nachteile.

Workarounds

Der DDI-Metadatenstandard DDI-L (vgl. Kapitel 9.1.3) hat zum Ziel, Daten und ihren Entstehungs- und Nutzungskontext entlang des von der DDI-Alliance entwickelten Forschungsdatenzyklus und seinen acht Phasen (vgl. Kapitel 2.2) zu dokumentieren. Dabei sollten die Metadaten diesen Prozess der Entstehung des Datensatzes und die zugrunde liegende Studie idealerweise auf Grundlage der konzipierten DDI-L Module darstellen. Der hier vorgestellte Workaround bricht insofern mit diesem Prinzip, indem die Geoinformationen nicht alleine auf Studienebene beschrieben werden. Vielmehr lassen sich mit dieser Lösung die geographischen Strukturen auch für einzelne Variablen auf der Ebene von Datensätzen beschreiben.

Dazu wird der georeferenzierte Umfragedatensatz, der mit räumlichen Geoinformationen aus drei verschiedenen Quellen verknüpft wurde, durch Nutzung des DDI-L-Moduls *LogicalProduct* in drei logisch voneinander getrennte Datensatz-Teile aufgegliedert. Jeder Teildatensatz enthält lediglich die Attribute, die aus den jeweiligen Geodatenquellen stammen. Darüber hinaus wird jeder der drei Teildatensätze durch eine *StudyUnit* beschrieben. Mittels dieser Vorgehensweise können nun Metadatenfelder implementiert werden, die zwar nur auf Studienebene anwendbar sind und somit für alle in dem jeweiligen Datensatz enthaltenen Variablen gelten. Da im entsprechenden Teildatensatz aber ohnehin nur Geoinformationen aus einer Quelle mit einer einzelnen geographischen Struktur enthalten sind, ist die Zuordnungsebene des Metadatenelements irrelevant.

Auch durch die Beschreibung auf Studienebene werden die aus den Geoinformationen stammenden Variablen akkurat beschrieben. Problematischer ist vielmehr die DDI-L-konforme Integration der Einzelstudien, die Teil eines Umfrageprogramms wie der Allgemeinen Bevölkerungsumfrage (ALLBUS) sind, wenn sie zuvor in einzelne logische Sinneinheiten geteilt wurden. Das kann jedoch wiederum mit gegenseitigen strukturierten Referenzierungen gelöst werden, indem das DDI-L-Modul *GroupPackage* genutzt wird, um Metadaten aus mehreren Einzelstudien zu verknüpfen.

Ein anderer Workaround besteht darin, dass gar nicht erst versucht wird, Geoinformationen zwingend in DDI zu beschreiben. Stattdessen werden separate Dateien mit Metadaten angelegt, die dann auf der Variablenebene in DDI referenziert werden. Das Format dieser Dateien könnte beispielsweise ebenfalls in einem XML-Format und somit strukturiert vorliegen.

Der größte Vorteil des ersten Workarounds ist, dass Metadaten wie üblich weiterhin in DDI beschrieben werden können. Gleichzeitig hat der Workaround aber auch den Nachteil, dass die Anzahl separater Metadatenobjekte schnell ansteigt, wenn viele verschiedene Geodatenquellen genutzt werden. Ähnlich verhält es sich auch mit dem zweiten Workaround. Zwar hat dieser Weg den Vorteil, dass das Prinzip der kompletten Beschreibung des Lebenszyklus von Forschungsdaten nicht aufgebrochen werden muss, gleichzeitig entstehen aber zwei Nachteile: Erstens steigt auch bei diesem Workaround die Anzahl separater Metadatenobjekte ggf. schnell an. Zweitens bleibt zwar die Kompatibilität mit DDI erhalten, die Informationen sind aber ggf. für auf dem DDI Standard basierende Software nicht mehr

zugänglich. Im Folgenden wird daher ein weiterer Weg der Implementierung vorgestellt. Dieser führt jedoch zu schematischer Invalidität mit DDI, die stets berücksichtigt werden muss.

Brute-Force-Ansätze

Ein radikaler Ansatz, ISO-19115-Metadaten auch auf der DDI-Variablenebene zu implementieren, besteht darin, diese Implementierung, obwohl konzeptuell nicht vorgesehen, zu erzwingen. Auch hier sind grundsätzlich zwei Wege denkbar:

Die Beschreibung geographischer Strukturen in DDI kann mittels des sogenannten *GeographicStructureScheme* und der darin enthaltenen hierarchisch untergeordneten Metadaten vorgenommen werden. Dies ist jedoch nur innerhalb der Studienebene möglich – auf Variableneben sind *GeographicStructureSchemes* nicht vorgesehen. Gleichzeitig handelt es sich bei DDI zunächst um reine Textdaten, die sehr wohl auf Variablenebene dem Standard widersprechend manipuliert werden können. Somit lassen sich einfache technische Routinen entwickeln, um die *GeographicStructureSchemes* auch auf der Variablenebene anwenden zu können.

Ein anderer und weniger invasiver Weg ist, auf der Studienebene mehrere *GeographicStructureSchemes* zu definieren. Die Idee ist dabei auch, einfache Referenzen innerhalb dieser *GeographicStructureSchemes* zu definieren, die auf die jeweilige Variable verweisen, deren Grundlage die jeweilige geographische Struktur ist. Dieser letzte Schritt ist in DDI jedoch bisher nicht vorgesehen. Daher muss hier wieder eine eigene – sprich neue – Metadatenstruktur implementiert werden, welches zu Inkompatibilitäten mit DDI führen kann.

Unabhängig davon, welcher Brute-Force-Ansatz gewählt wird, hat die Inkompatibilität mit DDI weitreichende Konsequenzen. In technische Systeme, die gemäß der DDI-L Spezifikation aufgebaut sind, können die in einem räumlichen Verknüpfungsprojekt erfassten Metadaten – gemäß des abweichenden Brute Force Ansatzes – nicht mehr umstandslos importiert und verarbeitet werden. Das ist vor allem dann problematisch, wenn z.B. das im Projekt bestehende Dokumentationssystem an erweiterte Katalogsysteme angeschlossen werden soll.

Die Dokumentation georeferenzierter Umfragedaten ist bislang wohl die größte Herausforderung, zumindest wenn diese auf standardisierten Metadaten beruhen soll. Der Einsatz von Workarounds oder Inkompatibilitäten mit dem DDI-Metadatenstandard erschweren die Dokumentation und erscheinen dadurch wenig attraktiv. Um diese Lücken angesichts der zunehmenden Relevanz von georeferenzierten Umfragedaten in der Forschung zu schließen, müssen entsprechende Metadatenstandards weiterentwickelt bzw. angepasst werden. Von dieser Problematik sind jedoch nicht allein georeferenzierte Umfragedaten betroffen, sondern auch andere bzw. neue Datentypen wie etwa die in Kapitel 11 vorgestellten *Social-Media*-Daten.

Entsprechende Entwicklungen im Bereich des DDI-L-Metadatenstandards sind bereits erkennbar angegangen worden. Das betrifft den Bereich der georeferenzierten Daten (Müller/Schweers/Zenk-Möltgen 2016; Müller/Schweers/Zenk-Möltgen 2015) ebenso wie von Social-Media-Daten (Borschewski/Zenk-Möltgen 2017). Trotzdem ist der gegenwärtige Zustand vage und ruft ggf. in Forschungsprojekten Unsicherheiten hervor. Dementsprechend muss auch der Forschungsgemeinschaft daran gelegen sein, Lösungen anzumahnen, sich an ihrer Entwicklung zu beteiligen und sie zu implementieren.

12.5 Weitergabe georeferenzierter Umfragedaten zur Sekundärnutzung

Bisher wurden vor allem Herausforderungen georeferenzierter Umfragedaten im Verlauf eines Forschungsprojekts erörtert. Zum Abschluss dieses Kapitels soll noch kurz auf den Umgang mit georeferenzierten Umfragedaten nach Projektende eingegangen werden. Im Kontext einer vorausschauenden Planung des projektinternen Forschungsdatenmanagements ist die primäre Herausforderung, einen rechts- und datenschutzkonformen Weg der Weitergabe zu finden, um so die Daten für die Sekundärnutzung verfügbar zu machen (vgl. Kapitel 3.2.2). Entsprechende Dienstleistungen von Datenarchiven, Datenzentren und Repositorien werden in Kapitel 7.4 thematisiert.

Das zentrale Problem der projektinternen Planung der Sicherung, Archivierung, Bereitstellung und Nachnutzung georeferenzierter Daten durch Dritte liegt im rechtskonformen Umgang mit Forschungsdaten und Materialien sowie dem datenschutzkonformen Umgang mit personenbezogenen Daten. Wie in Abschnitt 12.3 bereits erörtert, besteht bei georeferenzierten Umfragedaten durch den Raumbezug ebenso wie durch die Anreicherung mit Zusatzinformationen ein erhöhtes Re-Identifikationsrisiko befragter Personen. Diesem Risiko kann generell mit zwei Strategien begegnet werden: Zum einen können die Daten aggregiert werden, sodass eine Re-Identifikation faktisch ausgeschlossen ist. Zum anderen kann der Zugang zu den Daten kontrolliert und durch individuelle Nutzungsvereinbarungen geregelt werden.

Beide Ansätze haben Vor- und Nachteile. Das Aggregieren von Werten ist u.U. recht einfach zu bewerkstelligen und legt somit der Weitergabe etwa für Sekundäranalysen keine zusätzlichen Limitationen auf. Allerdings schränkt es – je nach Ausmaß der Aggregation – den analytischen Zugewinn ein und verringert das Analysepotential. Die Zugangskontrolle hingegen erhält zwar den analytischen Zugewinn, ist aber gleichzeitig mit erheblichem Mehraufwand in der Betreuung und Kontrolle der Nachnutzenden verbunden.

Die Weitergabe nicht-aggregierter Daten durch entsprechende Zugangskontrollen ist in der Regel nur durch Archive oder Repositorien und der Expertise ihrer Mitarbeitenden zu gewährleisten. Je nach Ausrichtung und Portfolio haben diese Distributionswege für besonders schützenswerte Daten entwickelt. So hat z.B. das GESIS Datenarchiv für Sozialwissenschaften mit dem *Secure Data Center* eine Einrichtung etabliert, die neben dem tatsächlichen Zugang zu den Daten und der anschließenden Kontrolle der Ergebnisse auch Beratungen im Bereich sensibler bzw. georeferenzierter Daten ermöglicht.

Durch entsprechende Einrichtungen können also mittels gesicherter Zugangswege die Vorzüge des *Data Sharings* genossen und dennoch eine Übereinstimmung mit datenschutzrechtlichen Fragestellungen gefunden werden. Und auch das gleichzeitige Angebot einer aggregierten Version der Daten ist damit nicht ausgeschlossen. Oft wird ein aggregierter und einfach zugänglicher Datensatz veröffentlicht (*Public Use File*) sowie ein weiterer besonders schützenswerter Datensatz parallel über gesicherte Zugangswege angeboten (*Scientific Use File*) (s. Kapitel 4.3.3).

Schließlich bleibt noch ein weiteres Kriterium zur Weitergabe georeferenzierter Umfragedaten zu berücksichtigen. Abhängig davon, mit welchen Geodaten georeferenzierte Umfragedaten räumlich verknüpft werden, können Rechte Dritter davon berührt sein. Das ist vor allem dann relevant, wenn Geoinformationen kommerzieller Anbieter verwendet werden. Aber auch Geodaten aus öffentlicher Hand könnten u.U. unter Nutzungsauflagen weitergegeben worden sein (Schweers et al. 2016: 110). Es sollte daher ebenfalls geprüft werden, ob entsprechende Vereinbarungen eine Weitergabe des georeferenzierten Datensatzes überhaupt erlauben. Fehlt diese Möglichkeit aufgrund mangelnder Urheber- bzw. Verwertungsrechte, bietet sich Forschenden die Alternative anstelle der Forschungsdaten die Skripte zu

deren Erstellung zu archivieren und anderen Forschenden verfügbar zu machen, wie in Kapitel 8.3 ausführlich diskutiert.

12.6 Zusammenfassung

Dieses Kapitel befasste sich mit den Möglichkeiten einer Georeferenzierung von Umfragedaten und deren räumlicher Verknüpfung mit Informationen aus Geodaten. Dabei wurden vor allem technische, organisatorische, datenschutzrechtliche und dokumentarische Herausforderungen, aber auch die Weitergabe georeferenzierter Umfragedaten erörtert. Zusammenfassend bleibt die Aussage, dass die Arbeit mit georeferenzierten Umfragedaten zugegebenermaßen anspruchsvoll ist.

Für jede der einzelnen Herausforderungen wurden verschiedene Antworten und Lösungen vorgestellt. So kann den technischen und organisatorischen Herausforderungen entweder mit eigener personeller Ressourcenplanung – entweder durch Weiterbildung oder gezielter Anwerbung – begegnet werden. Oder es wird die Expertise von Drittanbietern im Bereich der Georeferenzierung in Anspruch genommen. In Bezug auf die datenschutzrechtlichen Herausforderungen können Bedenken vor allem durch einen umsichtigen Umgang mit den Daten auf technischer sowie organisatorischer Seite begegnet werden. Eine Trennung von Adressdaten bzw. Geokoordinaten und den eigentlichen Umfragedaten ist dabei oberstes Prinzip, das sich in der physischen Speicherung der Daten selbst niederschlagen muss. Indessen finden sich bezüglich der Dokumentation räumlicher Verknüpfungen leider keine eindeutigen Antworten. Die vollständige Dokumentation ist zwar möglich, bedeutet aber je nachdem Workarounds oder Inkompatibilitäten mit dem für die Sozialwissenschaften relevanten DDI-L Standard. Im Zweifelsfall sollte hier auf die Beratung und Unterstützung durch disziplinäre Dienstleister, wie etwa dem GESIS Datenarchiv, zurückgegriffen werden.

Der Anwendungsfall georeferenzierter Umfragedaten und ihre Verknüpfung mit kleinräumigen Geodaten steht exemplarisch für eine Reihe von Innovationen und Trends in der empirischen Sozialforschung. So können einige in diesem Kapitel vorgestellte Herausforderungen durchaus auf die Verwendung anderer und neuer Datentypen wie etwa Social-Media- oder administrative Daten in den Sozialwissenschaften übertragen werden. Auch bei diesen Datentypen herrschen aktuell noch große Unsicherheiten etwa in Bezug auf deren Dokumentation (Borschewski/Zenk-Möltgen 2017). Sozialwissenschaftliche Forschung ist dynamisch. Entsprechend müssen die Prozesse im Forschungsdatenmanagement sowie relevante Metadatenstandard wie etwa DDI-L stets an neue Entwicklungen und Herausforderungen angepasst werden.

Literatur

AdV/KLA (2015): Leitlinien zur bundesweit einheitlichen Archivierung von Geobasisdaten. Abschlussbericht der gemeinsamen AdV-KLA-Arbeitsgruppe „Archivierung von Geobasisdaten“ 2014–2015. Hamburg. http://www.bundesarchiv.de/DE/Content/Downloads/KLA/leitlinien-geobasisdaten.pdf [Zugriff: 15.06.2018].

Blatt, Amy J. (2012): Ethics and Privacy Issues in the Use of GIS. In: Journal of Map & Geography Libraries 8, 1, S. 80-84. https://doi.org/10.1080/15420353.2011.627109.

Bluemke, Matthias/Resch, Bernd/Lechner, Clemens/Westerholt, René/Kolb, Jan-Philipp (2017): Integrating Geographic Information into Survey Research. Current Applications, Challenges and Future Avenues. In: Survey Research Methods 11, 3, S. 307-327. https://doi.org/10.18148/srm/2017.v11i3.6733.

Borschewski, Kerrin/Zenk-Möltgen, Wolfgang (2017): Facilitating Metadata Capture and Reuse in the Social Sciences with the Example of Social Media Data. Vortrag: 7th Conference of the European Survey Research Association (ESRA), Lissabon, 20.07.2017. https://www.europeansurveyresearch.org/conference/programme2017?sess=186&day=3 [Zugriff: 15.06.2018].

Dietz, Robert D. (2002): The Estimation of Neighborhood Effects in the Social Sciences. An Interdisciplinary Approach. In: Social Science Research 31, 4, S. 539-575. https://doi.org/10.1016/S0049-089X(02)00005-4.

Edwards, Paul N./Mayernik, Matthew S./Batcheller, Archer L./Bowker, Geoffrey C./Borgman, Christine L. (2011): Science Friction: Data, Metadata, and Collaboration. In: Social Studies of Science 41, 5, S. 667-690. https://doi.org/10.1177/0306312711413314.

El Emam, Khaled (2006): Overview of Factors Affecting the Risk of Re-identification in Canada (Access to Information and Privacy Division of Health Canada).

Esri (2015): ArcGIS Desktop: Release 10.3. Redlands, California: ESRI – Environmental Systems Research Institute.

Förster, André (2018): Ethnic Heterogeneity and Electoral Turnout. Evidence from Linking Neighbourhood Data with Individual Voter Data. In: Electoral Studies 53, S. 57-65. https://doi.org/10.1016/j.electstud.2018.03.002.

GESIS – Leibniz-Institut für Sozialwissenschaften (2015): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften ALLBUS 2014. GESIS Datenarchiv, Köln, ZA5240 Datenfile Version 2.1.0. https://doi.org/doi:10.4232/1.12288.

Goebel, Jan/Wagner, Gert G./Wurm, Michael (2010): Exemplarische Integration raumrelevanter Indikatoren auf Basis von „Fernerkundungsdaten" in das Sozio-oekonomische Panel (SOEP). SOEPpapers on Multidisciplinary Panel Data Research 267.

Hillmert, Steffen/Hartung, Andreas/Weßling, Katharina (2017): Dealing with Space and Place in Standard Survey Data. In: Survey Research Methods 11, 3, S. 267-287.

Klinger, Julia (2018): Allgemeine Bevölkerungsumfrage der Sozialwissenschaften – ALLBUS Sensitive Regionaldaten. GESIS Data Archive. https://doi.org/10.4232/1.13010.

Klinger, Julia/Müller, Stefan/Schaeffer, Merlin (2017): Der Halo-Effekt in einheimisch-homogenen Nachbarschaften: Steigert die ethnische Diversität angrenzender Nachbarschaften die Xenophobie? In: Zeitschrift für Soziologie 46, 6, S. 402-419. https://doi.org/10.1515/zfsoz-2017-1022.

Kounadi, Ourania/Leitner, Michael (2014): Why does Geoprivacy Matter? The Scientific Publication of Confidential Data Presented on Maps. In: Journal of Empirical Research on Human Research Ethics 9, 4, S. 34-45. https://doi.org/10.1177/1556264614544103.

Kroll, Martin/Schnell, Rainer (2016): Anonymisation of Geographical Distance Matrices via Lipschitz Embedding. In: International Journal of Health Geographics 15, 1, S. 1-14. https://doi.org/10.1186/s12942-015-0031-7.

Meyer, Reto/Bruderer Enzler, Heidi (2013): Geographic Information System (GIS) and its Application in the Social Sciences using the Example of the Swiss Environmental Survey. https://doi.org/10.12758/mda.2013.016.

Müller, Stefan/Schweers, Stefan/Siegers, Pascal (2017): Geocoding and Spatial Linking of Survey Data. An Introduction for Social Scientists. GESIS Paper 2017/15. S. 1-29. https://www.ssoar.info/ssoar/bitstream/handle/document/52316/ssoar-2017-muller_et_al-Geocoding_and_Spatial_Linking_of.pdf?sequence=1 [Zugriff: 15.06.2018].

Müller, Stefan/Schweers, Stefan/Zenk-Möltgen, Wolfgang (2015): Georeferenced Survey Data at the GESIS Data Archive. Vortrag: EDDI15 – 7th Annual European DDI User Conference, Copenhagen, 02.12.2015.

Müller, Stefan/Schweers, Stefan/Zenk-Möltgen, Wolfgang (2016): The Past, Present and Future of Geocoded Survey Data at the GESIS Data Archive. Vortrag: EDDI16 – 8th Annual European DDI User Conference, Cologne, 06.12.2016.

Plant, Richard E. (2012): Spatial Data Analysis in Ecology and Agriculture Using R. Boca Raton: CRC Press.

QGIS Development Team (2018): QGIS Geographic Information System. Open Source Geospatial Foundation Project. http://qgis.osgeo.org [Zugriff: 15.06.2018].

R Core Team (2017): R: A language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

RatSWD – Rat für Sozial- und Wirtschaftsdaten (2012): Endbericht der AG „Georeferenzierung von Daten" des RatSWD. Bericht der Arbeitsgruppe und Empfehlungen des RatSWD.

Saib, Mahdi-Salim/Caudeville, Julien/Carre, Florence/Ganry, Olivier/Trugeon, Alain/Cicolella, Andre (2014): Spatial Relationship Quantification Between Environmental, Socioeconomic and Health Data at Different Geographic Levels. In: International Journal of Environmental Research and Public Health 11, 4, S. 3765-3786. https://doi.org/10.3390/ijerph110403765.

Schweers, Stefan/Kinder-Kurlanda, Katharina/Müller, Stefan/Siegers, Pascal (2016): Conceptualizing a Spatial Data Infrastructure for the Social Sciences. An Example from Germany. In: Journal of Map & Geography Libraries 12, 1, S. 100-126. https://doi.org/10.1080/15420353.2015.1100152.

Skinner, Chris (2012): Statistical Disclosure Risk. Separating Potential and Harm. Statistical Disclosure Risk. In: International Statistical Review 80, 3, S. 349-368. https://doi.org/10.1111/j.1751-5823.2012.00194.x.
Stimson, Robert (2014): A Spatially Integrated Approach to Social Science Research. In: Stimson, Robert (Hrsg.): Handbook of Research Methods and Applications in Spatially Integrated Social Science. Cheltenham, United Kingdom: Edward Elgar, S. 13-25.
Strobl, Christian (2017): Dimensionally Extended Nine-Intersection Model (DE-9IM). In: Shekhar, Shashi/Xiong, Hui/Zhou, Xun (Hrsg.): Encyclopedia of GIS. New York, NY: Springer, S. 470-476.
Vardigan, Mary (2013): The DDI Matures. 1997 to the Present. In: IASSIST Quarterly 37, S. 45-50.
Vardigan, Mary/Heus, Pascal/Thomas, Wendy (2008): Data Documentation Initiative. Toward a Standard for the Social Sciences. In: International Journal of Digital Curation 3, 1, S. 107-113. https://doi.org/10.2218/ijdc.v3i1.45.
Weßling, Katarina D. (2016): The Influence of Socio-spatial Contexts on Transitions from School to Vocational and Academic Training in Germany. https://doi.org/10.15496/publikation-15222.
Zandbergen, Paul A. (2014): Ensuring Confidentiality of Geocoded Health Data. Assessing Geographic Masking Strategies for Individual-Level Data. In: Advances in Medicine 2014, S. 1-14. https://doi.org/10.1155/2014/567049.
Zenk-Möltgen, Wolfgang (2012): Metadaten und die Data Documentation Initiative (DDI). In: Altenhöner, Reinhard/Oellers, Claudia (Hrsg.): Langzeitarchivierung von Forschungsdaten. Standards und disziplinspezifische Lösungen. Berlin: Scivero, S. 111-126.

Linkverzeichnis

Bing: https://msdn.microsoft.com/de-de/library/ff701713.aspx [Zugriff: 15.06.2018].
BKG – Bundesamt für Kartographie und Geodäsie: https://www.geodatenzentrum.de/geodaten/gdz_rahmen.gdz_div/ [Zugriff: 15.06.2018].
DDI – Data Documentation Initiative: https://www.ddialliance.org/ [Zugriff: 15.06.2018].
Daten des Bundesamts für Kartographie und Geodäsie (BKG): http://www.geodatenzentrum.de/auftrag1/archiv/vektor/vg250_ebenen/ [Zugriff: 15.06.2018].
destatis – Statistisches Bundesamt: https://www.destatis.de/DE/ZahlenFakten/LaenderRegionen/Regionales/Gemeindeverzeichnis/NamensGrenzAenderung/NamensGrenzAenderung.html [Zugriff: 15.06.2018].
DSGVO – Datenschutzgrundverordnung: https://eur-lex.europa.eu/legal-content/DE/TXT/HTML/?uri=CELEX:02016R0679-20160504 [Zugriff: 15.06.2018].
DSAnpUG-EU – Datenschutz-Anpassungs- und -Umsetzungsgesetz EU: https://www.bmi.bund.de/SharedDocs/downloads/DE/gesetztestexte/datenschutzanpassungsumsetzungsgesetz.pdf [Zugriff: 15.06.2018].
EINONET Central Data Repository: https://cdr.eionet.europa.eu/ [Zugriff: 15.06.2018].
Geodateninfrastruktur Deutschland (GDI-DE): http://www.geoportal.de/ [Zugriff: 15.06.2018].
Geographic Structure Scheme: https://www.ddialliance.org/Specification/DDI-Lifecycle/3.2/XMLSchema/FieldLevelDocumentation/schemas/conceptualcomponent_xsd/elements/GeographicStructureScheme.html [Zugriff: 15.06.2018].
Google: https://developers.google.com/maps/documentation/geocoding/intro?hl=de [Zugriff: 15.06.2018].
GovData: https://www.govdata.de/ [Zugriff: 15.06.2018].
Infas 360: https://infas360.de/ [Zugriff: 15.06.2018].
INSPIRE: https://inspire.ec.europa.eu/ [Zugriff: 15.06.2018].
ISO-19115: https://www.iso.org/standard/73118.html [Zugriff: 15.06.2018].
microm: https://www.microm.de/ [Zugriff: 15.06.2018].
OpenStreetMap: https://www.openstreetmap.de/ [Zugriff: 15.06.2018].
OpenStreetMap Nominatim: https://nominatim.openstreetmap.org/ [Zugriff: 15.06.2018].
Secure Data Center:
https://www.gesis.org/angebot/daten-analysieren/weitere-sekundaerdaten/secure-data-center-sdc/ [Zugriff: 15.06.2018].
Sozio-oekonomisches Panel (SOEP): https://www.diw.de/de/soep [Zugriff: 15.06.2018].
Zensus 2011: https://www.zensus2011.de [Zugriff: 15.06.2018].