



# Daten, Metadaten, Interoperabilität

Herbert Kubicek, Andreas Breiter und Juliane Jarke

## Inhalt

1	Einführung .....	2
2	Begriffliche Grundlagen .....	2
3	Modellcharakter und Kontextabhängigkeit .....	5
4	Interoperabilität von Daten .....	7
5	Metadaten und Interoperabilität in der Praxis .....	9
6	Zusammenfassung und Schlussfolgerungen .....	12
	Literatur .....	13

## Zusammenfassung

Häufig ist von Daten die Rede, ohne dass der Terminus definiert wird. Es macht jedoch einen großen Unterschied, ob damit quantitative Angaben oder maschinenlesbare Zeichen gemeint sind. Auch wird meist nicht näher darauf eingegangen, dass der Umgang mit Daten im Wesentlichen auf Metadaten beruht und für den Austausch und die Zusammenführung unterschiedlicher Datenbestände Interoperabilität hergestellt werden muss. Dieser Artikel versucht, entsprechende Grundlagen zu schaffen. Zunächst werden die unterschiedlichen Daten-Begriffe gegenübergestellt. Dann wird auf den Modellcharakter und die Kontextabhängigkeit von Daten eingegangen, bevor Grundlagen der Interoperabilität und die Bedeutung von Metadaten in verschiedenen Anwendungsbereichen dargestellt werden. Abschließend wird kurz auf die wohl größte Herausforderung in diesem Zusammenhang eingegangen: auf die Qualität von Daten.

---

H. Kubicek (✉)

ifib: Institut für Informationsmanagement Bremen GmbH, Bremen, Deutschland

E-Mail: [kubicek@ifib.de](mailto:kubicek@ifib.de)

A. Breiter · J. Jarke

Universität Bremen und Institut für Informationsmanagement Bremen, Bremen, Deutschland

E-Mail: [abreiter@ifib.de](mailto:abreiter@ifib.de); [jjarke@ifib.de](mailto:jjarke@ifib.de)

---

**Schlüsselwörter**Big Data · Informationsbegriff · Semantik · Semiotik · Modelle

---

**1 Einführung**

Staat und Verwaltung sind ohne extensive Erhebung und Verarbeitung von Daten nicht denkbar. Ob jemand Bürgerin oder Bürger eines Staates ist, wird seit Jahrhunderten anhand von Daten dokumentiert, die in Registern gespeichert und in amtlichen Dokumenten ausgewiesen werden. Alle ordnungspolitischen Funktionen von Staat und Verwaltung basieren auf Daten und erzeugen Daten. Politische Entscheidungen und Gesetzesvorhaben beruhen auf der Analyse von Daten oder werden durch diese ausgelöst. Staat und Verwaltung nutzten von Anfang an die maschinelle Datenverarbeitung, *EDV (Elektronische Datenverarbeitung)*, *ADV (Automatisierte Datenverarbeitung)*, IT (Informationstechnik) genannt. Die erste umfassende Anwendung maschineller Datenverarbeitung war die Volkszählung in den USA im Jahr 1890. Das Urteil des Bundesverfassungsgerichts zur geplanten Volkszählung 1983 in Deutschland hat der staatlichen Datenverarbeitung durch die Prinzipien der Zweckbindung und Erforderlichkeit enge Grenzen gesetzt. Mit dem Internet wurde die bis dahin weitgehend verwaltungsinterne Datenverarbeitung sowohl zu anderen Verwaltungen als auch für Bürgerinnen und Bürger sowie die Wirtschaft geöffnet. Digitalisierung ist insofern eine Fortsetzung von sechzig Jahren Verwaltungsautomation. Mit *Open Data* soll nun nicht nur die Transparenz von Staat und Verwaltung verbessert, sondern auch wirtschaftliches Wachstum durch innovative digitale datengestützte Dienste gefördert werden. Und mit *Big Data* und Algorithmen aus dem Bereich der *Künstlichen Intelligenz* sollen neue und mächtigere Analyse- und Prognosemöglichkeiten entstehen. Stets hat die Erweiterung des Potenzials automatisierter Datenverarbeitung zu Visionen neuer Chancen und zu Befürchtungen neuer Risiken geführt. Kritiker überschätzen dabei häufig die Möglichkeiten der Nutzung und die Aussagekraft von Daten, Befürworter unterschätzen die hohen Voraussetzungen einer bereichsübergreifenden Realisierung der erhofften Wirkungen. Zu Fehleinschätzungen führen ein unscharfer Datenbegriff, eine mangelnde Berücksichtigung des Charakters von Daten als selektive Modelle sowie die unzureichende Auseinandersetzung mit der Bedeutung von Metadaten und der Interoperabilität von Daten aus unterschiedlichen Quellen.

---

**2 Begriffliche Grundlagen**

Das Wort „*Daten*“ hat in verschiedenen wissenschaftlichen Disziplinen, Politikfeldern und Verwaltungsbereichen unterschiedliche Bedeutung. In Lexika werden Daten definiert als quantitative Angaben oder durch Beobachtungen, Messungen, statistische Erhebungen gewonnene Angaben, meistens Zahlenwerte. Dies entspricht dem Verständnis von Daten in der Amtlichen Statistik. Etymologisch ist

das Wort eine Pluralbildung von „Datum“, lateinisch „gegeben“, (von *dare* = geben) bzw. als Substantiv „das Gegebene“. Darin liegt schon sprachlich eine Ambivalenz. Einerseits werden Daten als etwas Gegebenes angesehen, von dem man auszugehen hat: Die Daten einer Person (Name, Geburtsdatum etc.) oder die Rahmendaten in einer Volkswirtschaft. Gleichzeitig wird ausgedrückt, dass Daten gegeben, also erzeugt und zugeordnet werden, und zwar in einem sozialen Prozess. Volkswirtschaftliche Rahmendaten können durch staatliche Intervention und Gesetzgebung verändert werden.

Nach der Datenschutzgrundverordnung sind in Fortführung früherer Legaldefinitionen *personenbezogene Daten* nicht nur Zahlenwerte, sondern alle „Informationen, die sich auf eine identifizierte oder identifizierbare natürliche Person . . . beziehen“ (Artikel 4, Ziffer 1). Der Informationsbegriff bezieht sich nicht auf eine bestimmte materielle Form. Daten sind in diesem Sinne elektronische Daten in Datenbanken ebenso wie Akten, und auch Audio-, Bild- und Videoaufzeichnungen. Den Anlass für die Datenschutzgesetzgebung und das später im Volkszählungsurteil geschaffene *Recht auf informationelle Selbstbestimmung* bildeten die enorm gestiegenen Auswertungsmöglichkeiten elektronisch gespeicherter Daten, wie sie jetzt mit *Big Data* und maschinellem Lernen noch einmal eine neue Dimension erreichen. Im Volkszählungsurteil wurden der Verarbeitung personenbezogener Daten mit den Prinzipien der *Zweckbindung* und *Erforderlichkeit* deutliche Grenzen gesetzt. Es dürfen nur die Daten erhoben werden, die für einen definierten Zweck erforderlich sind, und sie dürfen auch nur für diesen Zweck verwendet werden (siehe auch den Artikel „Datenschutz“ in diesem Handbuch). 2001 wurde in einer Novellierung des Bundesdatenschutzgesetzes die Prinzipien „*Datenvermeidung* und *Datensparsamkeit*“ eingeführt, wonach die Erhebung, Verarbeitung und Nutzung von Daten und die Gestaltung von DV-Systemen an dem Ziel auszurichten sind, mit so wenig personenbezogenen Daten wie möglich auszukommen und wo immer möglich eine Pseudonymisierung oder Anonymisierung vorzusehen (§ 3a BDSG). In Artikel 5 der seit Mai 2018 geltenden Datenschutzgrundverordnung (DSGVO) finden sich das Zweckbindungsgebot in ähnlicher Formulierung und das Erforderlichkeitsprinzip als Grundsatz der „*Datenminimierung*“. Artikel 25 verlangt eine nach dem jeweiligen Stand der Technik ausgerichtete Gestaltung und datenschutzfreundliche Voreinstellungen, was durch Zertifizierung nachgewiesen werden kann. Diese Verpflichtung gilt für jede Art der Digitalisierung im öffentlichen Bereich.

In der *Informatik* wird unter Bezug auf die *Semiotik* zwischen Zeichen, Daten und Informationen unterschieden (z. B. Krcmar 2015, S. 11 ff.). Ein *Zeichen* ist etwas Sicht- oder Hörbares, mit dem jemand aufgrund einer Konvention auf etwas aufmerksam macht, z. B. eine arabische oder römische Ziffer oder ein *Bit* in einem digitalen System. Diese Zeichen entstammen einem gewachsenen oder vereinbarten Zeichenvorrat und können in unterschiedlichen Kulturräumen und Kontexten unterschiedliche Bedeutung haben. Das Zeichen X kann einen Buchstaben oder eine Zahl darstellen. Nach der semiotischen Theorie werden Zeichen nach bestimmten Regeln miteinander in Beziehung gesetzt, die Syntax oder Grammatik genannt werden. Aus dem Zeichenvorrat {0, 1, 2, 6, 8} kann eine Zeichenfolge im Format „12.06.2018“ oder „6.210,68“ gebildet werden.

Wofür die nach Syntaxregeln erzeugten Daten stehen, deren Bedeutung, nennt man *Semantik*. Die semantische Festlegung definiert die erste Zeichenfolge als Datum und die zweite als Währung, nicht als Länge einer Strecke. Daten sind in der Informatik jedoch eine besondere Art von Zeichen, nämlich solche, die maschineninterpretierbar sind. Dazu muss für jedes Datum die Syntax und Semantik festgelegt und kommuniziert werden. Die Syntax legt die Anzahl der Zeichen in einer Zeichenfolge und ihre Untergliederung durch Sonderzeichen (Punkte, Kommata) fest. Semantisch wird „12.06.2018“ als Datum in der deutschen Schreibweise und nicht in der englischen definiert, bei der der Monat vor dem Tag genannt wird. Und der Betrag „6.210,68“ als Währung in Euro in deutscher Schreibweise, mit der Trennung von Euro und Cent durch ein Komma. Es ist leicht erkennbar, dass es ohne diese Deklarationen zu Missverständnissen zwischen Computersystemen in Deutschland und England kommen kann. Man spricht in diesem Zusammenhang auch von einer Codierung der Daten und der Verwendung von Codes, wie einer Liste von Währungen, aber auch für Warengruppen und Gebiete nach Postleitzahlen, oder beim Amtlichen Gemeindegemeinschaftsschlüssel (AGS) des Statistischen Bundesamtes.

Im Kontext der Digitalisierung ist es wichtig, Daten von Texten klar abzugrenzen. In der *Open-Data*-Literatur wird pointiert unterschieden: *Texte* sind für Menschen verständlich, Daten für Maschinen. Texte sind *maschinenlesbar*, aber nicht maschinell interpretierbar und können deshalb nur bedingt weiterverarbeitet werden. Ein Textverarbeitungsprogramm kann jeden Text mit einer vereinbarten Zeichenmenge aus einem anderen Rechner syntaktisch korrekt wiedergeben, wenn das Format erkannt wird. Aber es kann die inhaltliche Bedeutung (Semantik) nicht ohne Weiteres erkennen. Mit Methoden der sogenannten Künstlichen Intelligenz (KI) sind neue Möglichkeiten entstanden, Texte zu verarbeiten, und weitere zu erwarten. So können auf Maschinellem Lernen basierende Spracherkennungsprogramme unter Berücksichtigung des Kontextes einzelner Wörter Texte zunehmend besser in eine andere Sprache übersetzen. Ob es sich bei dieser Interpretationsleistung um „Verstehen“ handelt, darüber streiten Sprachphilosophen, Hirnforscher und Kognitionsinformatiker seit Jahrzehnten.

Nach der Lehre der Semiotik unterscheiden sich Informationen von Daten durch ihren Zweck- und Handlungsbezug (Pragmatik). Dieselben Daten können für eine Person in einer bestimmten Situation relevantes handlungsbezogenes Wissen und damit eine wertvolle Information sein und für andere Personen völlig belanglos und uninteressant. Im ingenieurwissenschaftlichen Verständnis der ISO-Standards (*International Organization for Standardization*) sind Informationen hingegen generell Wissen, unabhängig davon, ob es handlungsrelevant ist. In der Betriebswirtschaftslehre wird beklagt, dass Unternehmen zu viele Daten, aber zu wenig Informationen haben. Es ist Aufgabe des *Informationsmanagements*, dies zu ändern, also Informationen zur richtigen Zeit, am richtigen Ort, in der richtigen Qualität bereitzustellen (Krcmar 2015). Für die Verwaltung gilt dies genauso. In beiden Bereichen werden *Big-Data*-Analyseverfahren als Mittel angesehen, um aus Daten Informationen zu machen.

Der Daten-Begriff in der internationalen technischen Normung bezieht sich nicht auf die Semiotik. Nach ISO/IEC 2382-1 „Informationstechnik – Begriffe“ (seit

1993) sind *Data*: „a reinterpretable representation of information in a formalized manner, suitable for communication, interpretation, or processing“ – eine wieder interpretierbare Darstellung von Information in formalisierter Art, geeignet zur Kommunikation, Interpretation oder Verarbeitung. Dabei wird der Begriff „Information“ nicht definiert.

In den Wirtschafts- und Rechtswissenschaften wird diskutiert, ob Daten *materielle* oder *immaterielle Güter* sind. Sicher ist, dass die physischen Objekte, auf denen Daten gespeichert werden (Datenträger), von den Daten selbst zu trennen sind. Strittig ist, ob und wie diese Daten wie Sachen zu behandeln sind, an denen man *Eigentum* erwerben kann, oder ob es sich um immaterielle, geistige Objekte handelt, an denen ein *Urheberrecht* besteht, aus dem die Überlassung von Nutzungsrechten abgeleitet werden kann. Da dies auch eine Rechtsunsicherheit in der Praxis zur Folge hat, wo Daten ein hohes wirtschaftliches Wachstumspotenzial zugeschrieben wird, wird über *Datengesetze* und neue Eigentumsformen *sui generis* diskutiert. Im Koalitionsvertrag von CDU, CSU und SPD von 2017 haben die Regierungsparteien angekündigt, die Frage, ob und wie ein *Eigentum an Daten* ausgestaltet sein kann, zügig anzugehen (Zeile 6085). Das Hauptproblem besteht darin, dass bei neueren digitalen Anwendungen die Erzeugung von Daten nicht mehr einem Urheber oder Eigentümer zugeordnet werden kann. Ein Beispiel sind die Bewegungsdaten von Kraftfahrzeugen, die durch ein vom Hersteller entwickeltes Programm auf einem vom Eigentümer mit dem Auto gekauften Datenträger gespeichert und nur durch das Fahrverhalten des Besitzers erzeugt werden. Ökonomen bezweifeln daher, dass Daten als Privatgüter anzusehen sind und schlagen stattdessen vor, sie als öffentliche Güter zu behandeln (Jentzsch 2018). In einem Gutachten für die Konrad-Adenauer-Stiftung stellt Karl-Heinz Fezer (2018) ein Eigentumsrecht *sui generis* des Bürgers oder Verbrauchers zur Diskussion, der durch sein Verhalten die Daten generiert und daher an daraus resultierenden Erlösen beteiligt werden soll.

Offen ist auch die Frage, ob Daten wie anderes Vermögen *vererbt* werden und ob Erben Eigentum an *online* gespeicherten Daten der Erblasser erwerben und daher den Zugang zu diesen Daten vom jeweiligen *Provider* verlangen können. Im Juli 2018 hat der Bundesgerichtshof entschieden, dass die Daten des *Facebook-Accounts* einer verstorbenen Tochter an die Eltern herauszugeben sind, weil es sich um eine Frage des Erbrechts handele und diese Daten auf einem externen Speicher genau so zu behandeln sind, wie ein Tagebuch im Zimmer der Tochter, das in das Eigentum der Erben übergeht. *Facebook* hatte dies mit Verweis auf seine Verpflichtung auf das Telekommunikationsgeheimnis zunächst verweigert.

---

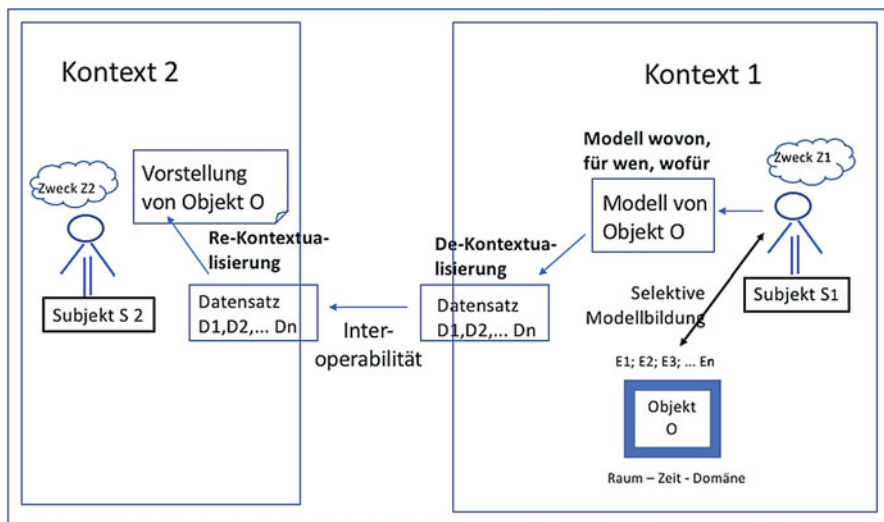
### 3 Modellcharakter und Kontextabhängigkeit

Die Unter- und Überschätzung von Daten resultiert aus einer ungenügenden Beschäftigung mit ihrer Selektivität und Kontextabhängigkeit. Diese ergibt sich aus dem Modellcharakter von Daten und hat Folgen für die Möglichkeit des Datenaustauschs zwischen verschiedenen verarbeitenden Stellen. Diese müssen durch Ver-

einbarungen zur Interoperabilität bilateral oder bereichsbezogen erst geschaffen werden (Abb. 1).

Zur Entwicklung von digitalen Anwendungen gehört die sogenannte Modellierung des Gegenstandsbereichs. Sie führt zu Datenmodellen, für die es verschiedene Modellierungssprachen gibt, wie z. B. die *Universal Modeling Language* (UML). Datenmodelle stellen dar, welche Daten mit welchen Beziehungen zueinander verarbeitet werden sollen. Sie sind eine Grundlage für die Anlage von Datenbanken. In der Informatik wird nicht immer deutlich gemacht, welche große Tragweite solche Modellierungsentscheidungen haben. Auch in der aktuellen Diskussion über *Big Data* wird mehr über die Algorithmen als über die Daten diskutiert, aus denen die Algorithmen Informationen machen (sollen).

In der Lehre der Semiotik stehen Zeichen in einer Dreiecksbeziehung zwischen einem Bezeichner (Subjekt) und etwas Bezeichnetem (Objekt). Diese bilden den Kontext. Nach Steinmüller (1993, S. 211 ff.) sind Daten Modelle von etwas für jemanden, die zu einem bestimmten Zweck geschaffen werden („Modell wovon – für wen – wofür“). Wie alle Modelle sind sie daher selektive Abbilder der jeweiligen Objekte. Einige Eigenschaften (E1, E2, ...) werden in das Modell aufgenommen und als Daten in einem Datensatz definiert (D1, D2, ...), aus dem dieser Kontext allerdings nicht mehr unmittelbar erkennbar ist, sie werden de-kontextualisiert. Über die Semantik wird dieser Kontext nur teilweise erfasst. Im Datenschutzrecht wird dieser Aspekt durch den Zweck der Verarbeitung erfasst und mündet in das Zweckbindungsgebot. Es ist verboten, Daten, die für einen bestimmten Zweck erhoben worden sind, ohne Erlaubnis für einen anderen Zweck – und damit in einem anderen Kontext – zu verwenden. Ein Beispiel ist die Wohnadresse einer Person, die als Anschrift angegeben wurde. Im Rahmen des sogenannten *Kredit-Scoring* werden



**Abb. 1** Daten als kontextabhängige Modelle und Interoperabilität (Quelle: Eigene Darstellung)

diese von Kreditauskunfteien aber auch als ein Indikator für die Kreditwürdigkeit der in einer bestimmten Nachbarschaft wohnenden Personen verwendet, also aus einem Kontext in einen anderen übertragen, obwohl nicht sicher ist, dass diese Daten im neuen Kontext dieselbe Validität haben. Mit der Neufassung des BDSG wurde in § 31 der Schutz der Betroffenen beim *Scoring* gestärkt. Anschriftendaten dürfen für die Wahrscheinlichkeitsberechnung nur verwendet werden, wenn die Betroffenen zuvor darüber unterrichtet wurden.

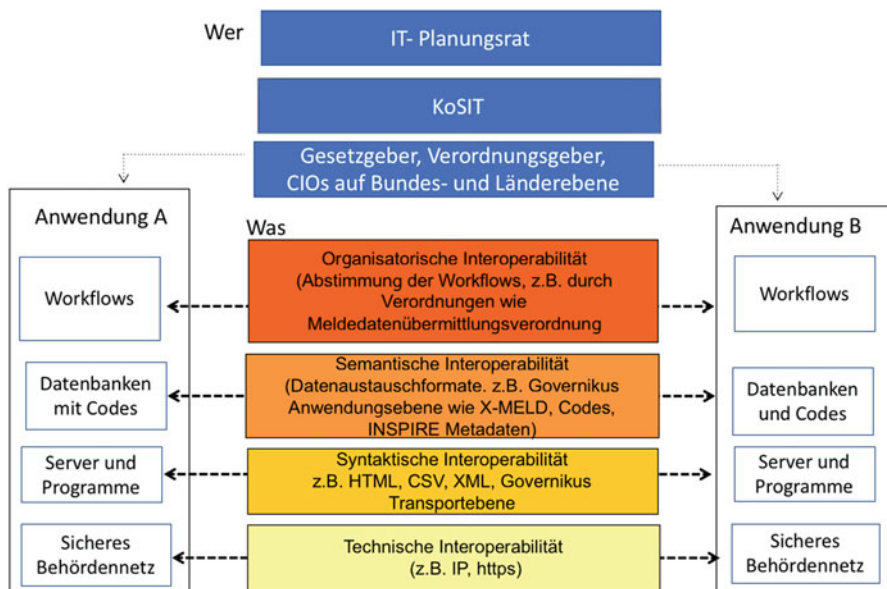
Die Selektivität, die Kontextabhängigkeit und der Zweckbezug von Daten („Modell wovon – für wen – wofür“) werden innerhalb der Sozialwissenschaften in den sog. *Critical Data Studies*, überwiegend im angelsächsischen Raum, thematisiert. Diese betonen, dass Daten nicht „gegeben“ sind, sondern „generiert“ werden und niemals „roh“ sind. Gitelman (2013) bezeichnet Rohdaten als *Oxymoron*, als einen Widerspruch in sich selbst. Daten sind nicht neutral oder objektiv, sondern existieren immer in Abhängigkeit von den Ideen, Instrumenten, Praktiken, Kontexten und dem Wissen, die sie generieren, verarbeiten und analysieren (Kitchin 2014). Sie sind keine rein technischen Artefakte, die einen Einfluss auf soziale Praxis haben, sondern werden innerhalb sozialer Praxis erzeugt, interpretiert und verarbeitet. Innerhalb der *Critical Data Studies* wird daher ein Schwerpunkt auf die Analyse der rekursiven Beziehung von Daten und Gesellschaft gelegt (Dalton et al. 2016; Jarke 2018), also danach gefragt, wie soziale Akteure Daten prägen und wie diese Daten wiederum soziale Wirklichkeit formen. Dabei wird ein weiterer Datenbegriff als der in der Informatik verwendet.

---

## 4 Interoperabilität von Daten

Wenn Daten im engeren Sinne von einer Quelle aus einem Kontext an eine andere Stelle zur maschinellen Weiterverarbeitung übermittelt werden, müssen die syntaktischen und semantischen Regeln vereinheitlicht, standardisiert werden. Es müssen Datenaustauschformate nach bestimmten Standards geschaffen werden. Sie machen unterschiedliche Quell- und Zielsysteme kompatibel oder interoperabel. Die Alternative wäre eine physische Zentralisierung der Daten in einer zentralen Datenbank, wie dies zum Beispiel Belgien und Österreich mit den Meldedaten in einem zentralen nationalen Register getan haben. Wenn man dies, wie in Deutschland, politisch nicht will, muss man mit rechtlichen Mitteln Interoperabilität zwischen den vielen hundert kommunalen Melderegistern herstellen. In diesem Fall wurde mit der Meldedatenübermittlungsverordnung der Standard X-MELD für den Austausch von Meldedaten im Rahmen der Rückmeldung zwischen der Wegzugs- und Zuzugskommune verpflichtend vorgeschrieben.

Im Einzelnen werden verschiedene *Ebenen der Interoperabilität* zwischen Systemen unterschieden, für die es unterschiedliche Arten von Standards gibt (Abb. 2). Auf der syntaktischen Ebene muss vereinbart werden, welche Zeichen und Formate verwendet werden und durch welche Zeichen einzelne Datenfelder voneinander abgegrenzt werden. Auf der semantischen Ebene muss vereinbart werden, aus welchen inhaltlichen Feldern ein Datensatz in welcher Reihenfolge besteht und



**Abb. 2** Ebenen der Interoperabilität und ihre Governance. (Quelle: Eigene Darstellung)

mit welchen Codes die Daten in den einzelnen Feldern erzeugt werden. Weiterverarbeitung erfolgt zu bestimmten Zwecken. Der pragmatische Aspekt wird durch die organisatorische Interoperabilität erfasst, indem die *Workflows* der Datenentstehung und der -verwendung aufeinander abgestimmt werden.

Davon zu unterscheiden ist die Organisation von Interoperabilität, auch als *Governance* bezeichnet, d. h. explizite Regelungen, wer Vorschläge wo einbringt, wer den Abstimmungsprozess zwischen den zu Beteiligten organisiert, wer die Anwendung der Standards verbindlich macht und wer die Einhaltung überwacht (ausführlicher Kubicek et al. 2011).

Für die Bundesverwaltung hat der IT-Beauftragte der Bundesregierung in Abstimmung mit den Ressorts einen SAGA (Standards und Architekturen für *E-Government-Anwendungen*) genannten Katalog der verpflichtend gemachten Standards zusammengestellt. Da viele *E-Government-Anwendungen* IT-Systeme von Bund, Ländern und Gemeinden betreffen, wurden 2009 mit der Föderalismusreform II die rechtlichen Voraussetzungen für eine lückenlose und medienbruchfreie elektronische Kommunikation zwischen den Behörden von Bund, Ländern und Kommunen geschaffen und mit dem Staatsvertrag über die Errichtung des *IT-Planungsrates* umgesetzt. Bund und Länder koordinieren in diesem Gremium ihre IT-Planung, verabschieden Standards und vereinbaren Pilotprojekte, bei denen jeweils ein Bundesland federführend ist (siehe den Artikel „IT-Planungsrat“ in diesem Handbuch). Für die Standardisierung gemeinsamer *E-Government-Dienste* wurde die *Koordinierungsstelle für IT-Standards* (KoSIT) gegründet, die u. a. auf Initiative der Fachministerkonferenzen tätig wird.



## 5 Metadaten und Interoperabilität in der Praxis

Interoperabilitätsstandards erlauben einen korrekten Austausch syntaktisch und semantisch definierter Daten für vereinbarte Zwecke. Sie erläutern jedoch nicht den Kontext, die Inhalte der Daten oder ihre Qualität. Dies ist Aufgabe sogenannter *Metadaten*: Daten über Daten. Auch hier gibt es zwei Ebenen: die inhaltliche Definition des Metadatensatzes und die *Governance* des Entwicklungs- und Verbreitungsprozesses. Denn Metadaten werden für Daten ganzer Bereiche oder Sektoren entwickelt, betreffen Daten vieler zumeist ähnlicher, aber geografisch verteilter Stellen und müssen daher auch interoperabel sein.

Der bekannteste Fall von Metadaten ist der *Bibliothekskatalog* (OPAC) mit Referenzdaten der Objekte im Bestand. Ein Metadatensatz beinhaltet hier Angaben zu Autor, Titel, Umfang, Verlag, ISSN oder ISBN und Standort sowie Zuordnung zu einem Themenbereich und passende Schlagwörter. Die beiden letzten Aspekte erweisen sich im Gegenstandsbereich der Bibliotheken, der alle Lebensbereiche und Wissensgebiete umfasst, als besonders schwierig für eine verbindliche Klassifizierung und Standardisierung. Für die Verschlagwortung ist ein *Thesaurus*, ein normierter und gepflegter Schlagwortbestand mit Synonymen und Homonymen, erforderlich. Die gut vernetzten deutschen Bibliotheken haben eine Schlagwortnormdatei als gemeinsamen Thesaurus geschaffen. Angesichts der *Online*-Kataloge im Internet wurde Mitte der 1990er-Jahre auf internationaler Ebene ein Metadatenstandard für Webressourcen entwickelt. Der Dublin Core Metadatensatz enthält 15 Felder zur einheitlichen Kennzeichnung von Ressourcen (<http://dublincore.org/>). Verantwortlich für Entwicklung, Weiterentwicklung und Pflege ist die *Dublin Core Metadata Initiative* (DSMI), die sich aus Freiwilligen aus Bibliotheken in der ganzen Welt zusammensetzt und in mehreren Arbeitsgruppen wirkt. Die Anwendung außerhalb von Bibliotheken ist gering. *Google* und andere Suchmaschinen haben den Bedarf nach einem derartigen offenen System aufgefangen.

Überwiegend sind Interoperabilitätsstandards auf der semantischen Ebene angesiedelt und Metadaten bereichsspezifisch. Am weitesten fortgeschritten ist die Standardisierung von Daten und Metadaten für *Umwelt- und Geodaten*. Hier ist die Abstimmungsnotwendigkeit besonders groß. Messwerte von Umweltbelastungen werden dezentral erfasst, sollen aber gemeinsam für größere Flächen bundesweit angezeigt und ausgewertet werden. Karten, für die Geo-Daten dezentral von Vermessungsämtern erhoben und in Katasterämtern aufbereitet und mit Inhalten gefüllt werden, sollen ebenfalls auf Landes- und Bundesebene aggregiert und mit weiteren Daten angereichert werden. Für Umweltdaten im Sinne der EU-Richtlinie über den Zugang der Öffentlichkeit zu Umweltinformationen (Richtlinie 2003/4/EG) wurde 2003 eine Vereinbarung zwischen Bund und Ländern geschlossen, die den gemeinsamen Betrieb sowie die Entwicklung und Pflege des Metainformationssystems *Umwelt-Datenkatalog* (UDK) und des Umweltinformationsnetzes Deutschland (*German Environmental Information Network* [GEIN]) regelt. Von 2006 bis 2014 hat das *PortalU* (wie „Umwelt“) die dezentralen Angebote der Verwaltungen nach den Umweltinformationsgesetzen der Länder und des Bundes unter einem gemeinsamen Dach zugänglich gemacht. Die inhaltliche und technische Verantwortung lag

bei der Koordinierungsstelle PortalU im Niedersächsischen Ministerium für Umwelt und Klimaschutz, die von Bund und Ländern gemeinsam finanziert wurde. Diese haben 2014 die Verwaltungsvereinbarung gekündigt, so dass das Portal ohne öffentliche Angabe von Gründen eingestellt wurde.

Für *Geo-Daten* gilt seit 2007 die EU-Richtlinie INSPIRE, die durch *Geodatenzugangsgesetze* auf Bundes- und Landesebene umgesetzt wurde und vorschreibt, wie durch Harmonisierung und Interoperabilität der Daten, Metadaten und Datendienste eine einheitliche Geodateninfrastruktur geschaffen wird. Auch hier haben Bund und Länder über eine Verwaltungsvereinbarung die organisatorischen Voraussetzungen für die Vernetzung von raumbezogenen Daten im Rahmen einer Geo-Daten-Infrastruktur Deutschland (GDI-DE) geschaffen. Betreiberorganisation ist die Koordinierungsstelle GDI-DE im Bundesamt für Kartografie und Geodäsie (BKG), die im Auftrag eines gemeinsamen Entscheidungs- und Steuerungsgremiums von Bund, Ländern und Kommunen tätig ist. In der Richtlinie wurden Fristen für die nationale Umsetzung einheitlicher Definitionen von genau definierten Geo-Basisdaten und Geo-Anwendungsdaten vorgegeben. 2016 wurde der letzte Umsetzungsbericht veröffentlicht.

Deutlich weniger standardisiert und durch behördliche Informationssysteme durchdrungen ist die *Bildungsverwaltung*. Erst durch Änderungen der Landes- und des Bundesstatistikgesetzes in den letzten fünf Jahren wird eine einheitliche Datenerlieferung eingefordert. Das gilt sowohl für die Schule (einheitliche Schüler-ID) als auch den Hochschulbereich (z. B. Erfassung von Promovierenden). Erst durch das Nationale Bildungspanel besteht die Möglichkeit zu Längsschnittstudien (<https://www.neps-data.de/>). Es gibt eine Oligopolstruktur bei den Softwareherstellern und wenig Interesse an länderübergreifender Standardisierung. Hierbei spielen landespolitische Aspekte (Kultushoheit) sowie datenschutzrechtliche Fragen eine Rolle. Zudem fehlen Abstimmungsgremien, in denen ein Aushandlungsprozess zwischen Herstellern, Verwaltung und Nutzern erfolgen könnte – weder die Kultusministerkonferenz noch die Gemeinsame Wissenschaftskonferenz sind hierzu hinreichend aufgestellt. Ein neues Feld, durch das eine stärkere Dynamik erwarten werden kann, sind *Offene Forschungsdaten*, die sowohl durch EU-weite Ansätze (*EU Science Cloud*) als auch Bund-Länder-Initiativen für Nationale Forschungsdateninfrastrukturen forciert werden.

Einerseits hoch standardisiert, andererseits stark individualisiert sind Gesundheitsdaten und damit auch die Gesundheitsverwaltung. Während medizinische Versorgungsdaten weltweiten Standards unterliegen und durch Krankenhausinformationssysteme und Systeme der Krankenkassen verarbeitet werden, sind die Schnittstellen zu anderen Versorgungseinrichtungen wie Apotheken, niedergelassenen Ärzten etc. kaum realisiert worden. Die Einführung der elektronischen Gesundheitskarte verzögert sich seit Jahrzehnten. Noch komplexer stellt sich die Situation für persönliche Daten der Bürger und Bürgerinnen dar, die im Zuge der „Selbstvermessung“ („Quantified Self“) gesammelt werden. Je nach *App* und Anbieter werden Daten nach unterschiedlichen Standards erhoben, ihre Kalibrierung wird systembezogen vorgenommen und Datenaustausch findet in der Regel nicht statt – teilweise

kann der Nutzer und die Benutzerin auf die eigenen Daten im geschlossenen Ökosystem eines Herstellers nicht mehr zugreifen (Lupton 2016).

Eine besondere Herausforderung an die Interoperabilität von Metadaten besteht bei den sogenannten *Offenen Verwaltungsdaten* (*Open Government Data*). Nach einer Verwaltungsvereinbarung zwischen dem Bund und einigen Ländern sollen alle Daten, die nicht speziellen Beschränkungen unterliegen, nicht nur auf ihren jeweiligen Landesportalen über Datenkataloge zur Weiternutzung bereitstellen, sondern die Metadaten auch an das Bundesportal *gov.data.de* gemeldet werden. Vier Länder, Bayern, Hessen, Niedersachsen und Sachsen-Anhalt, sind nicht beigetreten. Durch diese Zusammenführung sollen Entwickler, Wissenschaftler\*innen, Journalist\*innen und andere mit Hilfe eines zentralen Datenkatalogs erkennen, wo sie bestimmte Datensätze finden. Die Landesportale speisen sich wiederum teilweise aus kommunalen Datenportalen und erfassen diese Daten ebenfalls nicht selbst, sondern auch nur die Metadaten. Wie gut das Bundesportal Interessenten bei der Suche und dem Auffinden von Datensätzen unterstützt, hängt also entscheidend von diesen Metadaten ab. Diese sind aber auf der anderen Seite eine Hürde bei der Bereitstellung von Daten, weil sie von den datenhaltenden Stellen nach einheitlichen Regeln definiert und eingegeben werden müssen. Bisher gibt es unterschiedliche Metadaten-Standards für Offene Verwaltungsdaten. Neben dem von der *Open Knowledge Foundation* in Großbritannien (OKF) entwickelten Datenkatalog CKAN (*Comprehensive Knowledge Archive Network*) mit einem eigenen Metadatensatz, der von vielen kommunalen Portalen verwendet wird, hat der IT-Planungsrat nach einem ersten Standardisierungsversuch mit einem eigenen nationalen Metadatensatz 2018 einen für Bund und Länder verbindlichen Metadatenstandard verabschiedet. Dabei handelt es sich um eine deutsche Adaption des EU-Standards DCAT-AP, der das bisherige Metadatenmodell OGD.1.1 ablösen soll (*Data Catalogue Application Profile* des *W3C Data Catalog* DCAT). Als Vorteile nennt das für die Erarbeitung dieses Vorschlags federführende Portal Sachsen einen noch einfacheren Einstieg in die Veröffentlichung durch wenige Pflichtfelder, die Einführung der Kategorie „empfohlene“ Informationen als Kompromiss zwischen Verbindlichkeit und Flexibilität, die Existenz einer ganzen Reihe verbindlicher Vokabulare und die Interoperabilität zu anderen europäischen Portalen. Mit nur fünf Angaben als Pflichtfelder in diesem Metadatensatz ist die Aussagekraft für Interessenten sehr begrenzt. Generell fehlt auch im Vergleich zu anderen Bereichen eine Verpflichtung zur Einhaltung selbst dieser minimalen Metadaten.

Besondere politische Beachtung findet die Interoperabilität von Daten bei den verschiedenen Sicherheitsbehörden auf Länder- und Bundesebene sowie auf internationaler Ebene. Immer wieder wird nach Berichten über Straftaten, die bei einem Datenaustausch zwischen Behörden eventuell hätten vermieden werden können, eine bessere Abstimmung und teilweise auch eine Zusammenführung von Dateien gefordert und häufig wohl auch vorgenommen. Einzelfälle belegen, dass die interoperable Vernetzung zwar Möglichkeiten schafft, aber noch nicht garantiert, dass auch alle vorgesehenen Daten tatsächlich weitergegeben werden.

Aktuell stehen die Verwaltungen von Kommunen, Landes- und Bundesbehörden mit dem „Once Only-Prinzip“ vor der wahrscheinlich größten datenbezogenen

Verwaltungsreform, deren Bedeutung jedoch noch nicht allen Beteiligten deutlich zu sein scheint. *Once Only* heißt, die Verwaltung darf nur einmal nach bestimmten Daten fragen und muss, bevor sie Bürger\*innen oder Unternehmen fragt, prüfen, ob die begehrten Daten nicht bereits bei einer anderen Stelle vorliegen. Um Millionen von Anfragen zwischen Behörden zu vermeiden, läuft das darauf hinaus, für bestimmte Daten nur noch ein zentrales Register einzurichten. In Belgien wurde dazu ein Gesetz erlassen, das für definierte Daten landesweit eine einzige authentische Quelle (*one authentic source*) vorschreibt. In anderen Ländern heißen sie Basisregister. Dabei geht es nicht nur um Personenstammdaten und Melderegister. In der Gesamtheit der Implementierungen in der EU gibt es weitere zentrale Register für Daten über ausländische Einwohner, Personenstand, Unternehmen, Liegenschaften, Einkommen, Bezug von Sozialleistungen, die den von Bürgern und Unternehmen zu leistenden bürokratischen Aufwand erheblich reduzieren können. Während auf EU-Ebene das *Once Only*-Prinzip ein Grundprinzip des E-Government-Aktionsplans 2016 bis 2020 ist und Mitgliedstaaten wie Belgien, Estland oder die Niederlande die gesetzlichen, technischen und organisatorischen Voraussetzungen für die Umsetzung geschaffen haben, gibt es in Deutschland nur vage Absichtserklärungen des IT-Planungsrats. Ein zentrales Bundesmelderegister wurde bisher von den Bundesländern abgelehnt. Nach wie vor muss man bei jedem Antrag bei jeder Behörde immer wieder dieselben Angaben zur Person und Adresse machen, obwohl sie nach dem erwähnten X-MELD-Datensatz einfach abgefragt werden könnten. Das E-Government-Gesetz hat dazu die datenschutzrechtlichen Voraussetzungen geschaffen, doch der Widerstand gegen die Aufgabe eigener Datenbestände ist groß.

---

## 6 Zusammenfassung und Schlussfolgerungen

Die Kontextabhängigkeit von Daten hat nicht nur eine semantische Dimension, sondern vor allem eine pragmatische, auf die Verwendung bezogene. In verschiedenen Handlungskontexten, Verwaltungsbereichen und Politikfeldern werden ganz unterschiedliche Anforderungen an die Qualität von Daten sowie die Garantie und Haftung für diese Qualität gestellt. An die Korrektheit und Aktualität von Fahndungsdaten, Lebensmittelwarnungen oder Messungen von Luftbelastungen in Verbindung mit Fahrverboten werden andere Anforderungen gestellt als an Daten der Amtlichen Statistik und wieder andere als an einen Veranstaltungskalender auf einem kommunalen Internetportal. Datenhaltende Stellen können in manchen Fällen, wie bei den Offenen Daten, die Gewährleistung ausschließen, in anderen Fällen nicht. Der kurze Überblick über die Interoperabilität von Daten zeigt auch, dass es große bereichsspezifische Unterschiede in den Bemühungen um eine hohe Datenqualität gibt.

Für die aktuelle Diskussion zu Digitalisierung, *Big Data* und *Open Data* sind daraus mindestens zwei Schlussfolgerungen zu ziehen:

- Pauschale Aussagen über die Bedeutung oder die Folgen von *Big Data* und *Open Data*, wie sie häufig getroffen werden, sind unseriös und irreführend, wenn

einzelne Beispiele verallgemeinert werden. Erforderlich ist eine Differenzierung nach Herkunfts- und Verwendungskontexten, um das Potenzial und dessen Realisierungsmöglichkeiten verlässlich einschätzen zu können.

- Es fehlen Beispiele für eine angebliche datengetriebene („data driven“) Digitalisierung bzw. Verwaltungsreform. Die hier geschilderten Beispiele zeigen eher, dass Verwaltungen in vielen Bereichen sich schwertun, die erforderlichen Daten zusammenzutreiben.

---

## Literatur

- Dalton, Craig M., Linnet Taylor, und Jim Thatcher. 2016. Critical data studies. A dialog on data and space. *Big Data & Society* 3(1). <https://doi.org/10.1177/2053951716648346>.
- Fezer, Karl-Heinz. 2018. *Repräsentatives Dateneigentum. Ein zivilgesellschaftliches Bürgerrecht*. Sankt Augustin/Berlin: Konrad Adenauer Stiftung. [https://www.kas.de/c/document\\_library/get\\_file?uuid=f828a351-a2f6-11c1-b720-1aa08eacff9&groupId=252038](https://www.kas.de/c/document_library/get_file?uuid=f828a351-a2f6-11c1-b720-1aa08eacff9&groupId=252038). Zugegriffen am 07.10.2019.
- Gitelman, Lisa, Hrsg. 2013. „Raw Data“ Is an Oxymoron. Cambridge, MA: MIT Press.
- Jarke, Juliane. 2018. Digitalisierung und Gesellschaft. *Soziologische Revue* 41(1): 3–20.
- Jentzsch, Nicola. 2018. *Dateneigentum – Eine gute Idee für die Datenökonomie?* Berlin: Stiftung Neue Verantwortung e. V. [https://www.stiftung-nv.de/sites/default/files/nicola\\_jentzsch\\_dateneigentum.pdf](https://www.stiftung-nv.de/sites/default/files/nicola_jentzsch_dateneigentum.pdf). Zugegriffen am 07.10.2019.
- Kitchin, Rob. 2014. *The data revolution. Big data, open data, data infrastructures and their consequences*. Los Angeles: Sage.
- Krcmar, Helmut. 2015. *Informationsmanagement*, 6. Aufl. Berlin/Heidelberg: Springer Gabler.
- Kubicek, Herbert, Ralf Cimander, und Hans Jochen Scholl. 2011. *Organizational Interoperability in E-Government. Lessons from 77 European Good-Practice Cases*. Heidelberg: Springer.
- Lupton, Deborah. 2016. *The quantified self*. Cambridge: Polity.
- Steinmüller, Wilhelm. 1993. *Informationstechnologie und Gesellschaft. Einführung in die Angewandte Informatik*. Darmstadt: Wissenschaftliche Buchgesellschaft.