



Reinhold Stahl
Patricia Staab

Die Vermessung des Datenuniversums

Datenintegration mithilfe
des Statistikstandards SDMX

 Springer Vieweg



Die Vermessung des Datenuniversums

Reinhold Stahl
Patricia Staab

Die Vermessung des Datenuniversums

Datenintegration mithilfe des
Statistikstandards SDMX

Reinhold Stahl
Dornburg
Deutschland

Patricia Staab
Frankfurt
Deutschland

ISBN 978-3-662-54737-3

ISBN 978-3-662-54738-0 (eBook)

DOI 10.1007/978-3-662-54738-0

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

Springer Vieweg

© Springer-Verlag GmbH Deutschland 2017

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften. Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag, noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Planung: Dr. Annika Denkert

Gedruckt auf säurefreiem und chlorfrei gebleichtem Papier

Springer Vieweg ist Teil von Springer Nature

Die eingetragene Gesellschaft ist Springer-Verlag GmbH Deutschland

Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

Über dieses Buch

Dieses Buch richtet sich an alle, die mit Daten arbeiten – sie sammeln, integrieren und auswerten müssen oder sie analysieren möchten – und dazu die Hilfsmittel der Informationstechnologie einsetzen. Es ist von Insidern geschrieben, gewissermaßen von Datenexperten für Datenexperten. Denn unser Geschäft ist die industrielle Erhebung und Verarbeitung großer Datenmengen zur Informationsgewinnung und die Bereitstellung der Ergebnisse als Basis für wichtige Entscheidungen.

Den überwiegenden Teil unserer beruflichen Erfahrungen haben wir im Bereich der statistischen Informationssysteme der Deutschen Bundesbank gesammelt. Das in diesem Buch präsentierte Gedankengut bezieht sich jedoch auf die globale Rolle der Statistik beim Aufbau umfassender Datenwelten und bei der Bereitstellung von Information als öffentliches Gut, nicht auf die speziellen Merkmale der Notenbankstatistiken. Deshalb stammen die im Folgenden aufgeführten Beispiele mehrheitlich auch bewusst nicht aus der Welt der Finanzwirtschaft oder einer anderen wissenschaftlichen Spezialdisziplin. Wir wollen anhand von Beispielen mit Alltagsrelevanz zeigen, wie die Konzepte der Statistik die Basis für eine universelle und standardisierte Bereitstellung von beliebigen Informationen bieten und sozusagen einen *barcode of information* liefern können, wie er zum Beispiel aus der Warenwirtschaft bekannt ist. Wir möchten also zeigen, dass für beliebige Informations- und Datenpunkte, nehmen wir als Beispiel die „durchschnittliche Schneehöhe in Garmisch-Partenkirchen im Januar 2016“, genauso eine Identifikation gebildet werden kann wie für „einen Liter Fair-Trade-H-Milch, Fettanteil 1,8 %“.

Für alle Informationsdienstleister stellt sich seit Jahren immer wieder die Herausforderung, explosionsartig wachsende Datenwelten erforschbar und nutzbar zu machen. Dies ist aber nicht nur ein Problem großer Datenmengen und zahlreicher Quellen. Tatsächlich liegt in der korrekten Verknüpfung von Daten unterschiedlicher Quellen und Themengebiete sowohl die größte Herausforderung als auch das größte Potenzial zum Aufbau von Wissen. Dieses Vorgehen wurde schon früh in der Kriminalistik eingesetzt und mit dem negativ belegten Begriff der „Rasterfahndung“ assoziiert. Inzwischen wird in nahezu allen Unternehmen und Institutionen versucht, unter dem – inzwischen positiv belegten – Begriff der „Datenintegration“ den Wissensaufbau durch Zusammenführung von Informationen aus unterschiedlichen Geschäftsbereichen oder Wissenschaftsdisziplinen voranzutreiben.

Die Vision, die wir in diesem Buch beschreiben, besteht in einer standardisierten Datenwelt, vergleichbar mit einem Baukastensystem, also einem „Lego für Daten“. Darin liegen Daten zu unterschiedlichsten Themengebieten in einer leicht zugänglichen und zueinander passenden Form vor. Diese Datenwelt kann kontinuierlich durch Einfügung neuer Inhalte erweitert werden. Je nach Erfordernis kann sie zur Problemerkennung und -lösung sowie zum Wissensaufbau verwendet werden, indem die Inhalte (Bausteine) flexibel miteinander verknüpft und inhaltlich und technisch zu neuen „Informationsgebilden“ geformt werden.

Es geht also um eine geordnete Sammlung relevanten Wissens, ein Datenkompendium, das gewissermaßen als „öffentliches Gut“ genutzt werden kann. Dabei mag sich der Begriff „öffentlich“ auf ein spezifisches Unternehmen, eine Branche, eine Wissenschaftsdisziplin oder – noch globaler gedacht – auf eine buchstäblich öffentliche Kollektion aus geografischen, meteorologischen, medizinischen, finanziellen, verkehrstechnischen, landwirtschaftlichen, versorgungstechnischen, pädagogischen und psychologischen Datenbeständen beziehen, möglicherweise ergänzt um Registerdaten wie Kataster- oder Unternehmensverzeichnisse (sofern zugänglich).

Wir glauben, dass sich diese Vision nur mit der disziplinierten Einhaltung bewährter Prinzipien und unter Nutzung neuester Methoden und Technologien verwirklichen lässt. Die von uns präferierten Ansätze stellen wir in unserem Buch vor. Dabei gehen wir besonders auf die beiden für die Datenarbeit relevanten Erfolgsfaktoren ein: die Einführung eines universell nutzbaren Ordnungssystems verbunden mit dem gemeinsamen Willen zur Standardisierung. Wir stellen den in der internationalen öffentlichen Statistik genutzten Standard *SDMX (Statistical Data and Metadata Exchange)* vor und zeigen auf, welche tiefgreifenden Veränderungen durch die Einführung dieses Standards und des damit verbundenen Ordnungssystems für die Arbeit der internationalen Statistikcommunity möglich waren. Wir glauben, dass der Schritt zur Standardisierung der Durchbruch zur gewinnbringenden Nutzung großer Datenmengen für vielfältige Themen ist. Und dass daher die Nutzung eines weltweit etablierten ISO-Standards wie *SDMX* auch für andere Themenbereiche der entscheidende Erfolgsfaktor sein kann. Dabei sind stets die ersten Schritte eines Standards von der Idee einiger weniger bis zum marktbeherrschenden Selbstläufer die schwersten. Auch diese beschreiben wir in den folgenden Kapiteln.

Die Motivation zum Schreiben dieses Buches besteht in diesem Werben für Standardisierung, für ein universelles Ordnungssystem für Daten und für die Verwendung der Konzepte und Standards der Statistik zum Aufbau dieser Datenwelt. Wir möchten damit alle die erreichen, die zum Ziel der Information als öffentliches Gut beitragen können, von der Wissenschaft über die Softwareindustrie bis hin zu intensiv datennutzenden Unternehmen und Institutionen, von der Leitungsebene bis hin zur Arbeitsebene. Wir richten uns daher sowohl an den Professor als auch an den wissenschaftlichen Mitarbeiter, sowohl an den Softwarearchitekten als auch an den Programmierer, sowohl an den *CIO (Chief Information Officer)* als auch an den *Data Analyst*. Dies soll aber nicht bedeuten, dass wir unseren Leserkreis auf die genannten Rollen beschränkt sehen möchten. Vielmehr schließen wir alle diejenigen mit ein, die sich für unser Thema interessieren.

Dieses Buch ist bewusst kein Fachbuch, kein „SDMX für Einsteiger oder Experten“ geworden, es soll vielmehr eine Hinführung zum Aufbau eines Ordnungssystems für die Datenwelten leisten. Andererseits dürfte es ohne eine konkrete Vorstellung davon, wie SDMX aufgebaut ist, schwer zu vermitteln sein, warum ausgerechnet dieser Standard das Potenzial für die Umsetzung unserer oben genannten Vision hat. Deshalb wird in diesem Buch nach den grundsätzlichen Ausführungen von Teil 1 auch eine gut verständliche, aber dennoch mit der erforderlichen Detailtiefe ausgestattete Beschreibung des Standards SDMX in Teil 2 erfolgen.

Wir hoffen, dass unsere Gedanken inspirierend und hilfreich sind, nicht zuletzt, weil ein Leser sich darin vielleicht wiedererkennen kann. Und wir möchten mit unseren Überlegungen dazu einladen, in der aktuellen *Big-Data-Bewegung* eine andere Perspektive einzunehmen, statt immer weniger über Daten nachzudenken und immer mehr auf die Schlagkraft der IT-Systeme zu bauen. Bei allem technischen Fortschritt stehen nach unserer Ansicht immer noch die Intelligenz, der Ideenreichtum und die Erfahrung des Menschen, der die Technik nutzt, im Vordergrund. Denn auch in „*Think Big*“ steht immer noch das „*Think*“ an erster Stelle.

In diesem Buch vertreten wir unsere persönlichen, aus unserer langjährigen Praxis und aus dem intensiven Gedankenaustausch mit Kollegen von anderen Organisationen, Herstellern, Software- und Beratungshäusern gewonnenen Ansichten. Diese spiegeln selbstverständlich nicht zwangsläufig die Ansicht der Deutschen Bundesbank oder ihrer Mitarbeiterinnen und Mitarbeiter wider.

Frankfurt am Main, Februar 2017

Reinhold Stahl, Dr. Patricia Staab

Über die Autoren

Die Autoren des Buches verfügen über langjährige praktische Erfahrung in der IT-technischen Realisierung statistischer und datenanalytischer Anforderungen. Sie verfügen daher über das für diese Schnittstellenarbeit typische Kompetenzprofil, nämlich die Kombination aus fachspezifischem Wissen über die verwendeten Daten, solider Grundlage an mathematisch-statistischer Methodik und angewandter Expertise in Softwareengineering



Reinhold Stahl Diplom-Mathematiker, ist seit 1985 im Statistikbereich der Deutschen Bundesbank beschäftigt. Dort baute er zunächst das statistische Informationsmanagement in der heutigen Form auf, bevor er 2014 die Stelle als Leiter der Statistik antrat. Die Erfolgsgeschichte des in diesem Buch vorgestellten SDMX-Standards begleitete er seit den Anfängen aktiv mit und führte diesen Standard für die Bundesbank-Statistik ein. Die Möglichkeiten, die sich durch die Standardisierung eröffneten, machten ihn zu einem überzeugten Befürworter dieser Vorgehensweise.



Dr. Patricia Staab promovierte Mathematikerin, begann im Jahr 2000 in der Bundesbank-Statistik ihre Arbeit – am Aufbau eines hausinternen statistischen Informationssystems, das auf dem SDMX-Standard basierte. Seit dieser Zeit haben sich sowohl der Standard als auch die statistischen Informationssysteme der Bundesbank, die sie zurzeit verantwortet, stark weiterentwickelt – der prägende Eindruck von der Schlagkraft des Standards ist jedoch geblieben.

Inhaltsverzeichnis

Teil I Mit Standardisierung zur umfassenden Datenwelt

1 Ausgangslage, Vision und Wegbeschreibung	3
1.1 Explodierende Datenwelten.....	3
1.2 Unzugängliche Datensilos.....	4
1.3 Der Kick liegt in der Verknüpfung.....	5
1.4 Die Verknüpfung gelingt mit einem Ordnungssystem.....	6
1.5 Das Ordnungssystem SDMX.....	7
2 Wie sieht die Realität aus?	11
2.1 Lücken trotz Sammelwut.....	11
2.2 Fehlende Ordnung im Datenuniversum.....	12
2.3 Nutzung der IT-Technologie nicht ohne fachliche Expertise möglich.....	12
Literatur.....	13
3 Was können wir von Big Data erwarten?	15
3.1 Der Big-Data-Hype.....	15
3.2 Was ist Big Data? Eine technische Betrachtung.....	16
3.3 Was leistet Big Data nicht?.....	17
3.4 Ethische Bedenken.....	19
3.5 SDMX und Big Data: Ergänzung statt Widerspruch.....	20
Literatur.....	22
4 Warum ist Datenintegration so schwierig?	23
4.1 Was ist Datenintegration?.....	23
4.2 Schnelligkeit der Entwicklung in der Informationstechnologie.....	26
4.3 Konkurrenzsituation von IT-Anbietern und Produkten.....	27
4.4 IT-Projekte statt Fachprojekte.....	27
4.5 Mentalität des Individualismus.....	28
4.6 Silodenken vor fachübergreifendem Denken.....	29
4.7 Datenschutz.....	30
4.8 Fehlende unmittelbare Anreize für Datenanbieter.....	31

4.9	Ungenügende informationstechnische Standards für Daten	32
	Literatur	33
5	Grundsätzliche Einschätzung der Standardisierung.	35
5.1	Standards fallen nicht vom Himmel	35
5.2	Standards sind nirgends optimal, wohl aber das Optimum	36
5.3	Standards setzen sich dann durch, wenn sie nutzbar sind	36
5.4	Standards fördern dezentrales Arbeiten	37
5.5	Standards zur Verwirklichung völlig neuer Ansätze – aktuelles Beispiel: Blockchain	37
6	Forschung und Standardisierung	41
6.1	Begrenzttes Interesse an Standardisierung	41
6.2	Einfluss des Datenmaterials auf die Forschung	41
6.3	Rolle der Forschungsdatenzentren (FDZ)	42
	Literatur	44
7	Standards erfolgreich einführen	45
7.1	Die richtige Reihenfolge – der inhaltliche Einstieg	45
7.2	Struktur und Ordnung schaffen	46
7.3	Klassifizierungssysteme und Schlüssel nutzen	47
7.4	Technik richtig einsetzen	47
7.5	Die richtige Schrittlänge wählen	48
7.6	Stakeholder richtig behandeln	49
8	Statistik als Treiber erfolgreicher Datenintegration	51
8.1	Statistik als fachübergreifend generische Disziplin	51
8.2	Konzepte der Statistik zum Aufbau einer Datenwelt	52
8.3	Datenaustausch und Data Sharing in der Statistik	53
	Literatur	54
9	Beitrag des Statistikstandards SDMX	55
9.1	Was ist SDMX?	55
9.2	Einstieg in SDMX	56
9.3	SDMX an einem vereinfachten Beispiel	57
9.4	Data Driven Systems im Statistikdatenaustausch dank SDMX	59
9.5	Ausgereiftes Beispiel aus der Praxis	60
	Literatur	62
10	Fazit und Ausblick.	65
	Literatur	66
Teil II Der Statistikstandard SDMX		
11	Entstehung und Entwicklung von SDMX.	69
11.1	Die Idee, ihre Entstehung und Ausbreitung	69
11.2	Der Weg zum weltweiten Standard: Die SDMX-Initiative	71

11.3 Die Weiterentwicklung durch die Gremien der SDMX-Initiative	73
11.4 Das Potenzial: Nutzung als Information Model	76
11.5 Die Zukunft: Weitere Nutzungsmöglichkeiten, stärkere Industrialisierung	77
Literatur	78
12 Die wesentlichen Elemente von SDMX	79
12.1 Grundbausteine	79
12.2 Eine Datenstruktur wird definiert	80
12.3 Die Struktur wird mit Daten gefüllt, es entsteht ein Datensatz	83
12.4 Datensätze werden versandt und ausgetauscht	84
12.5 Die größere Perspektive – Verwaltung von Informationen, Themenbereichen, Akteuren, Prozessen	88
12.6 Das SDMX-basierte Data Warehouse	90
12.7 Anwendbarkeit von SDMX für Mikrodaten	91
12.8 SDMX und benachbarte Standards	92
Literatur	94
13 Arbeiten mit SDMX	95
Literatur	97
14 SDMX als Erfolgsfaktor für eine gelungene Datenintegration	99
Literatur	100
Glossar	101
Weiterführende Literatur	103
Stichwortverzeichnis	105