# Metadaten

## – Eine Grundlage für die Auswertung amtlicher Statistiken durch die Wissenschaft –

**Von Alexander Richter und Dr. Stefan Weil**

Erschließt sich dem Angehörigen unseres Kulturkreises die Interpretation des Kalenderdatums noch intuitiv – der 28.2.2005 ist nun einmal der 28. Tag im Februar dieses Jahres und als ein Montag auch gleichzeitig der erste Tag in der 9. Kalenderwoche – ,ist hingegen für eine korrekte Interpretation statistischer Daten ein Mehr an Informationen nötig. Um Daten der amtlichen Statistik exakt auswerten zu können, bedarf es der Angaben darüber, wie, wann, warum oder auch durch wen diese Daten erhoben wurden. Der erste Teil des Beitrages veranschaulicht Sinn und Zweck solcher „Daten über Daten". Im zweiten Teil wird eine Möglichkeit des Zugangs zu diesen so genannten Metadaten am Beispiel einer gemeinsamen Metadatenbank der Forschungsdatenzentren der statistischen Ämter des Bundes und der Länder (FDZ) dargestellt.

### Was sind Daten – was sind Metadaten?

In einer Welt, in der die Information einen fast schon dominierenden Produktionsfaktor darstellt, haben wir es ständig mit Daten zu tun. Tagtäglich nehmen wir eine Unzahl davon auf, interpretieren sie, speichern sie – im Gehirn, auf Papier oder auf einem elektronischen, magnetischen oder optischen Medium. In nicht allzu ferner Zukunft wird vielleicht sogar ein Ablegen solcher Informationseinheiten auf tesafilm® tägliche Praxis sein!

Statistische Ämter erheben Daten in einer fast unüberschaubaren Vielfalt. Das Internet und andere zumeist elektronische Medien helfen, diese Daten auf schnellstem Wege an die Stellen zu verbreiten, wo sie nachgefragt und genutzt (also weiterverarbeitet) werden. Auf ihrer Grundlage werden dann vielleicht wichtige und weit tragende Entscheidungen getroffen.

Trotz dieser mittlerweile herausragenden Bedeutung von Daten machen wir uns in der Regel keine Gedanken darüber, was ein Datum eigentlich ist. Es ist sicherlich nicht abwegig zu behaupten, dass kaum jemand in der Lage ist, den Begriff treffend zu definieren. Der Griff zum Wörterbuch oder Lexikon ist vorprogrammiert. Der Perfektionist wird einen Schritt weiter gehen und (beispielsweise) die einschlägigen Normen des Deutschen Instituts für Normung (DIN) oder der International Organization for Standardization (ISO) konsultieren. Im Dokument zur ISO/IEC-Norm 11179-3 wird er in der Tat folgende Definition vorfinden: „data[:] a re-interpretable representation of information in a formalized manner suitable for communication, interpretation or processing".[1)]

1) International Organization for Standardization (ISO): Information technology – Metadata registries (MDR) – Part 3: Registry metamodel and basic attributes. International Standard ISO/IEC 11179-3. Second edition, 2003-02-15. Genf 2003, S. 6.

Daten repräsentieren Informationen

Dass es sich bei einem Datum um eine Repräsentation, gewissermaßen um ein Muster einer Information handelt, erscheint uns dabei nachvollziehbar, vielleicht sogar trivial. Wesentlich sind jedoch die Reinterpretierbarkeit und die Darstellungsform eines Datums. Beide machen Daten für Kommunikation, Interpretation und Verarbeitung erst tauglich. Was ist damit gemeint?

Daten müssen eindeutig (re-)interpretierbar sein

Ein Datum muss eindeutig der Informationseinheit zuzuordnen sein, die es repräsentiert, und zwar dergestalt, dass nicht nur ein und dieselbe Person oder ein und dieselbe Maschine (Computer) zu unterschiedlichen Zeiten und/oder an unterschiedlichen Orten die Zuordnung fehlerfrei und kongruent sicherstellt. Vielmehr müssen auch unterschiedliche Personen und unterschiedliche Maschinen das Datum stets fehlerfrei und kongruent der repräsentierten Einheit zuordnen können.

Ein Beispiel soll dies verdeutlichen. Hierzu ziehen wir das (Kalender-)Datum, gewissermaßen als begriffliche „Mutter aller Daten“, heran. Das Datum „28.2.2005“ muss, soll es für alle dieselbe Bedeutung haben, eindeutig interpretierbar sein, und zwar als der 28. Tag des Monats Februar im Jahr 2005 nach Christi Geburt.[2)] Der im wahrsten Sinne tägliche Umgang mit einem Datum lässt uns vergessen, dass diese Interpretation im Sinne der obigen Definition durchaus nicht trivial ist. Diese Interpretation ist nämlich, wenn auch vielleicht nicht offiziell, doch zumindest faktisch, auf den Kulturkreis der westlichen Welt beschränkt, wo der gregorianische Kalender verwendet wird. In der islamischen Welt, wo der islamische Kalender Anwendung findet[3)], wird diesem Tag als Datum der 19.1.1429[4)] (das ist der 19. Tag des Monats Muharram des Jahres 1429 nach der Flucht Mohammeds von Mekka nach Medina)[5)] zugeordnet. Umgekehrt entspräche das islamische Datum 28.2.2005 dem christlichen Datum 11.1.2567.

Metadaten sind Voraussetzung für die eindeutige Reinterpretierbarkeit von Daten

Für die eindeutige (Re-)Interpretation des Datums benötigen wir somit weitere Informationen, also Informationen über das Datum; wir können auch sagen: Daten zum Datum. Solcherlei Daten über Daten bezeichnet man als „Metadaten“ (von griechisch meta = mit ..., nach ..., zwischen ...). Metadaten beinhalten gewissermaßen „Hintergrundinformationen“, welche die vielfältigen Eigenschaften von zumeist umfangreichen und komplexen Datenbeständen beschreiben und dadurch den inhaltlichen Kontext herstellen (Beschreibungs- und Erklärungsfunktion). Im gewählten Beispiel wäre ein für die richtige Interpretation des Kalenderdatums notwendiges Metadatum eine Information darüber, welcher Kalender Verwendung findet.

## Metadaten sind in der amtlichen Statistik unerlässlich

Eine verbindliche Definition des Begriffes „Metadaten“ findet sich in der ISO-Spezifi-

2) Dabei sei an dieser Stelle einmal angenommen, es bestünde Einigkeit über das tatsächliche Geburtsjahr Jesu (Konvention).

3) Der islamische Kalender wurde im Jahr 638 n. Chr. von Umar, dem zweiten Kalifen des Islam, eingeführt. Im Gegensatz zum gregorianischen Kalender basiert er auf dem Mondjahr, das aus zwölf Monaten besteht. Ein Mondmonat dauert ca. 29,5 Tage. Damit hat das Jahr des islamischen Kalenders 354 (genauer gesagt: 29,5 x 12) Tage. Als einziges dem islamischen Kulturkreis zugehöriges Land hat die Türkei (1927) den gregorianischen Kalender übernommen. In der Tat ist jedoch – vor allem im Wirtschaftsleben – auch in der islamischen Welt der gregorianische Kalender in Gebrauch.

4) Umgerechnet mit dem Konvertierungs-Tool auf http://www.ori.unizh.ch/hegira.html [Stand: 13.12.04] (für die Richtigkeit übernehmen die Autoren keine Gewähr).

5) Der erste Tag des islamischen Kalenders entspricht dem 16. 7. 622 im gregorianischen Kalender.

kation 11179-3. Hier heißt es sinngemäß, dass es sich dabei um Daten handelt, die der Definition und Beschreibung anderer Daten dienen.[6] Übertragen auf die Zwecke der amtlichen Statistik heißt dies also: Metadaten sind all diejenigen Informationen, die für die korrekte Interpretation von statistischen Daten notwendig sind. Die amerikanische Statistik definiert „statistische Metadaten" sogar noch umfassender: „Statistical metadata is descriptive information or documentation about statistical data, i.e. microdata, macrodata, or other metadata. Statistical metadata facilitates sharing, quering, and understanding of statistical data over the lifetime of the data."[7]

Metadaten steigern die Effizienz der Datenverarbeitung und -nutzung

Obwohl, wie aus dieser Definition hervorgeht, Metadaten auch Metadaten erklären können, sollten solche Datenbeschreibungen keiner weiteren Erläuterung bedürfen. Insofern sind statistische Metadaten Beschreibungen, die gewährleisten, dass die in den Datenbeständen (Datensätzen) der amtlichen Statistik enthaltenen Informationen für alle Anwender über einen möglichst langen Zeitraum gleichermaßen verständlich und interpretierbar sind. Diese – und ähnliche – Definitionen erklären Metadaten primär aus der Sicht des Nutzers von Daten der amtlichen Statistik. Eine vollständige Beschreibung des Begriffs sollte zusätzlich auch die Sicht der Produzenten und Verarbeiter umfassen. Ein Beispiel hierfür ist die Definition von Bethlehem u. a. Sie subsumieren unter Metadaten all diejenigen Informationen, die relevant und erforderlich sind, um Daten zu sammeln, zu verarbeiten und im weitesten Sinne zu nutzen.[8] Sie weisen so darauf hin, dass Metadaten in allen Produktions-, Verarbeitungs- und Nutzungsprozessen von Relevanz sind. Diese Definition gewährleistet damit eine effiziente Kommunikation zwischen allen Instanzen, die sich mit Daten der amtlichen Statistik befassen. Das sind die Auskunftspflichtigen (sie müssen die Fragebogen verstehen, um sie korrekt ausfüllen zu können), die Statistiker (sie erheben und verarbeiten die Daten) wie auch die Wissenschaftler (ihnen muss es möglich sein, die gewünschten Daten zu finden, auszuwählen, zu verstehen und zu bewerten sowie zielführend zu verwerten).[9]

Metadaten gewährleisten somit einen intelligenten und effizienten Zugriff auf die Daten[10], und Datenbestände können über längere Zeiträume hinweg personenunabhängig gepflegt werden.

Metadaten sind Bestandteil eines Metadatenschemas

Werden Metadaten systematisch verwaltet, sind diese in der Regel integraler Bestandteil eines Metadatenschemas, in welchem die Struktur der Metadaten beschrieben wird. Dieses Metadatenschema ist wiederum Teil eines Metadatenmodells. Die Metadaten werden üblicherweise in einer Metadatenbank oder einem Metadatenregister hinterlegt.

6) International Organization for Standardization (ISO): Information technology – Metadata registries (MDR) – Part 3: Registry metamodel and basic attributes. International Standard ISO/IEC 11179-3. Second edition, 2003-02-15. Genf 2003, S. 7.

7) Dippo, C.S./Gillman, D.W.: The Role of Metadata in Statistics. Statistical Commission and Economic Commission for Europe Working Paper No.2, Genf, September 1999, S. 1.

8) Bethlehem, J. et al.: On the Use of Metadata in Statistical Data Processing. Statistical Commission and Economic Commission for Europe Working Paper No. 23, Genf, September 1999, S. 3.

9) Vgl. ebenda.

10) Vgl. Marugg, T.: Wissens-Management: Metadaten für Content-Indizierung und Wissenssicherung, Teil 1, online im Internet: http://www.internetmanagement.ch/index.cfm/fuseaction/shownews/newsid/351 [Stand: 14.12.04]. Dabei bedient man sich metadatenbasierter, standardisierter Abfragesprachen, wie zum Beispiel der Structured Query Language (SQL), die sich zur Standardsprache für den Zugriff auf relationale Datenbanken entwickelt hat.

## Die Metadatenbank der Forschungsdatenzentren der statistischen Ämter

Amtliche Statistik ermöglicht den Zugang zu Mikrodaten

Die statistischen Ämter des Bundes und der Länder haben es sich zur Aufgabe gemacht, der Wissenschaft ausgewählte, faktisch anonymisierte[11] Mikrodaten[12] für Analysen zur Verfügung zu stellen. Der Aufbau von Forschungsdatenzentren auf Bundes- und auf Länderebene dient dem Ziel, anerkannten Forschungsinstituten und Universitäten den Zugang zu Einzeldaten aus verschiedensten Erhebungsbereichen der amtlichen Statistik zu ermöglichen. Als Dienstleister tragen die statistischen Ämter somit einem immer weiter steigenden Bedarf an Informationen Rechnung, welche die Grundlage für nachhaltige politische, ökonomische und gesellschaftliche Entscheidungen bilden.

Ein vereinfachter Zugang der Wissenschaft zu den Daten ist allerdings nur ein Schritt auf dem Weg zur Auswertung von Erhebungen der amtlichen Statistik. Um die Mikrodaten korrekt interpretieren zu können, bedarf es, wie oben bereits angedeutet, der Metadaten.

Metadaten verhindern Fehlinterpretationen

Selbst bei auf den ersten Blick relativ selbsterklärenden amtlichen Erhebungen wie der Geburtenstatistik können fehlende Informationen zu Fehlinterpretationen führen. Während sich bei der Auswertung des Merkmals „Geschlecht des Kindes“ auch ohne zusätzliche Informationen durch Metadaten kaum Schwierigkeiten ergeben dürften, stellt sich dies bei näherer Untersuchung des Merkmals „Lebend- oder Totgeburt“, vor allem bei der Betrachtung über einen längeren Zeitraum, völlig anders dar. Eine Geburt gilt in Deutschland als Lebendgeburt, wenn nach der Trennung vom Mutterleib entweder die Nabelschnur pulsiert oder der Herzschlag oder die Lungenatmung eingesetzt hat. Totgeborene sind dagegen Kinder, bei denen bei einem Geburtsgewicht von mindestens 500 g keines dieser drei Merkmale in Erscheinung tritt. Liegt das Geburtsgewicht unter 500 g, handelt es sich um eine Fehlgeburt.

Bis zum 31. März 1994 war für diese Abgrenzung ein Geburtsgewicht von 1000 g entscheidend, und bis zum 30. Juni 1979 galt eine Körperlänge von mindestens 35 cm als maßgeblich für die Unterscheidung einer Tot- von einer Fehlgeburt.[13]

Da Fehlgeburten in den Personenstandsbüchern nicht beurkundet werden und somit auch nicht in der Geburtenstatistik erscheinen, können allein solche Definitionsänderungen dazu führen, dass sich die Zahl an (Tot-)Geborenen nach bestimmten Stichtagen ändert. In diesem Fall kann ein Mangel an Metainformationen zu einer Fehlinterpretation statistischer Daten durch den Nutzer führen.

Die Geburtenstatistik und das darin enthaltene Merkmal „Lebend- oder Totgeburt“ stehen an dieser Stelle nur stellvertretend für eine Vielzahl von Statistiken und Merkmalen, die der Wissenschaft während und nach der Aufbauphase der Forschungsdatenzentren der statistischen Ämter des Bundes und der Länder zur Verfügung ste-

11) Weiterführende Informationen zur faktischen Anonymität von Mikrodaten finden sich online im Internet: http://www.forschungsdatenzentrum.de/anonymisierung.asp [Stand: 31.12.04].

12) Mikrodaten sind Angaben zu einzelnen Personen, Unternehmen oder sonstigen Einheiten.

13) Siehe hierzu Statistisches Bundesamt (Hrsg.): Fachserie 1: Bevölkerung und Erwerbstätigkeit, Reihe 1: Gebiet und Bevölkerung 2000-2002, S. 7, 2004.

hen werden. Das Beispiel aber zeigt, dass es unerlässlich ist, dem Nutzer der FDZ nicht nur den Zugang zu Mikrodaten, sondern auch zu weiterführenden Informationen über diese Daten zu ermöglichen.

Metadaten umfassen nicht nur Merkmalsdefinitionen

Für den adäquaten Umgang mit den Mikrodaten der amtlichen Statistik sind neben Erklärungen zur Merkmalsdefinition auch Informationen zu den Ausprägungen einzelner Merkmale, zu den rechtlichen Grundlagen der Erhebung oder Angaben über den Erhebungszeitpunkt und -umfang Voraussetzung.

Metadatenbank erleichtert den Zugang

Damit den Nutzern der Forschungsdatenzentren diese Fülle an Informationen in angemessener Weise zur Verfügung gestellt werden kann, ist die Einrichtung einer Metadatenbank unabdingbar. Aus diesem Grund wird parallel zu den Datenbeständen der Forschungsdatenzentren derzeit auch ein Metadatensystem aufgebaut. Als Basis für dieses System dient die bereits erprobte und über das Internet zugängliche Technologie von GENESIS[14], dem gemeinsamen Informationssystem der statistischen Ämter des Bundes und der Länder. GENESIS ermöglicht es dem Nutzer, sich schnell und bequem einen Zugang zu den bereits bestehenden Datenangeboten der amtlichen Statistik zu verschaffen. Im Gegensatz zu den der Wissenschaft in den Forschungsdatenzentren zur Verfügung gestellten Daten gestattet GENESIS lediglich den Zugriff auf Makrodaten, also Daten, die in aggregierter Form vorliegen, und den dazugehörigen Metadaten.

Für die speziellen Zwecke der Forschungsdatenzentren werden die einzelnen Komponenten der GENESIS-Technologie erweitert und um neue Funktionen ergänzt. Durch die Nutzung bereits bewährter Datenbankstrukturen wird ein rationeller Umgang mit den materiellen und personellen Ressourcen der amtlichen Statistik gewährleistet. Der Zugang zu den erweiterten Funktionen wird zukünftig ebenfalls über das Internet möglich sein.

Zwei Wege zur Recherche: hierarchisches Schlüsselsystem oder stichwortorientierte Suche

Zur Recherche von Informationen werden zwei Wege angeboten. Die erste Variante sieht eine hierarchische Erschließung der Metadaten vor. Die amtlichen Statistiken sind nach dem einheitlichen System EVAS[15] verschlüsselt. Auf der obersten Ebene wählt der Wissenschaftler den EVAS-Einsteller aus. Beim eingangs beschriebenen Beispiel der Geburtenstatistik wäre der entsprechende EVAS-Einsteller die „1“ für den Bereich der Bevölkerungsstatistiken. Auf der zweiten Ebene entscheidet sich der Nutzer mit Hilfe eines fünfstelligen Schlüssels für eine dem Bereich zugeordnete Statistik, z. B. „12612“ für die Geburtenstatistik.

Die zweite Variante basiert auf einer stichwortorientierten Suche, bei der ein entsprechender Suchbegriff zu den Statistiken verweist, in denen der gesuchte Begriff vorhanden ist. Handelt es sich bei dem gesuchten Wort beispielsweise um eine Merkmalsausprägung, werden dem Wissenschaftler alle entsprechenden Merkmale, in denen dieser Begriff auftaucht, mit Hinweis auf die jeweiligen Statistiken ausgegeben. Die Eingabe des Wortes „weiblich“ führt den Nutzer somit zum einen zu unserem Beispiel der Geburtenstatistik (das hierin vor-

14) Gemeinsames Neues Statistisches Informations-System.

15) Einheitliches Verzeichnis aller Statistiken.

kommende Merkmal „Geschlecht des Kindes“ verfügt schließlich über diese Ausprägung), zum anderen aber auch zu weiteren Statistiken, bei denen Angaben zum Geschlecht erhoben wurden. Über die so ermittelten Statistiken kann sich der Nutzer die Metadaten zu den einzelnen Erhebungen anzeigen bzw. ausdrucken lassen.

Für die Nutzung der Metadatenbank der Forschungsdatenzentren gilt es zwischen einer Statistik und einer Erhebung zu unterscheiden. Die Statistik entspricht dem oben erwähnten EVAS-Fünfsteller (z. B. 12612 Geburtenstatistik). Eine Erhebung ist die zu einem bestimmten Zeitpunkt bzw. für eine bestimmte Periode getätigte Durchführung einer Statistik (z. B. Erhebung aller Geburten im Jahr 2004).

Die Metainformationen über die Mikrodaten sind in der Datenbank vier grundlegenden Bereichen zugeordnet:

Metadaten umfassen vier Bereiche: ...

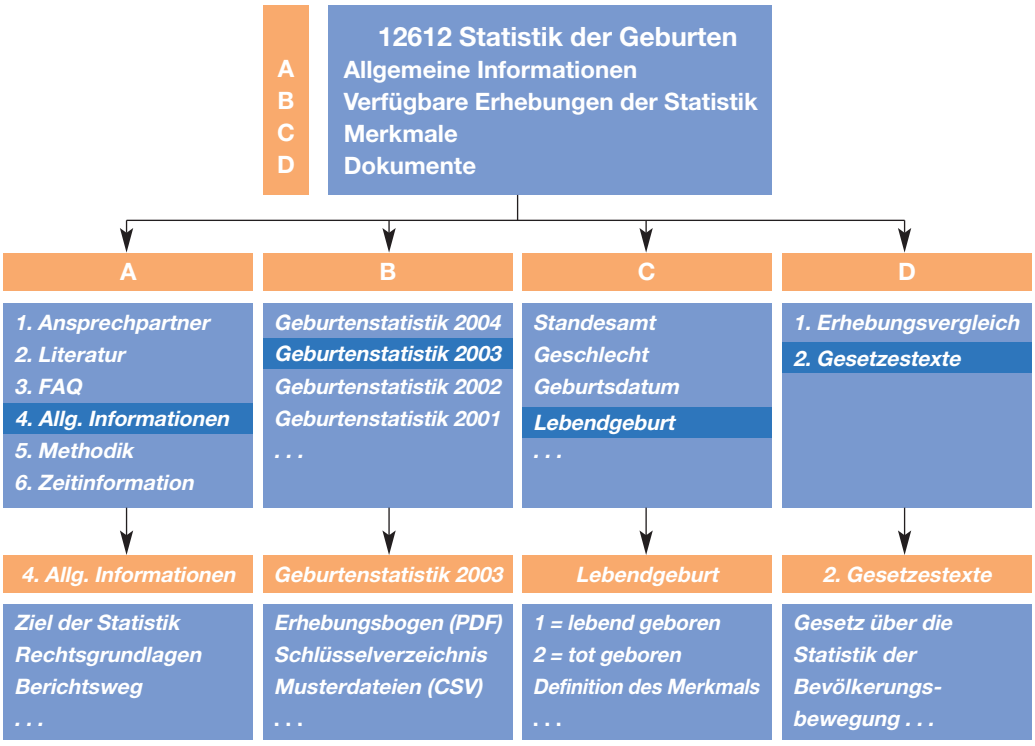
Im Bereich A kann sich der Wissenschaftler über allgemeine Angaben zu den einzelnen Statistiken informieren. Der Nutzer erfährt hier beispielsweise etwas über Ansprechpartner in den statistischen Ämtern, über Methodik und Periodizität der Erhebung sowie über weiterführende Literatur zum Themengebiet. Zusätzlich erhält der Wissenschaftler Informationen zu Aufbau, Berichtsweg und regionaler Tiefe der Statistik.

... Angaben zur Statistik ...

Der Bereich B dient der Angabe über die von den Forschungsdatenzentren zur Verfügung gestellten Erhebungen einer Statistik, für die der Nutzer einen Zugang beantragen kann. Hier finden sich zusätzlich

... Angaben zur Erhebung ...

**S 1** **Schema zum Aufbau der Metadatenbank**



erhebungsspezifische Informationen und Dokumente wie Fragebogen, Schlüsselverzeichnisse, Qualitätsberichte usw., die dem Wissenschaftler in vielen Fällen als Datei zum Download bereitstehen.

... Definition von Merkmalen ...

Bereich C bezieht sich auf die Definitionen von Merkmalen einer Erhebung und deren Ausprägungen. Für das Merkmal „Lebend- oder Totgeburt“ bekäme der Nutzer zu der bereits oben beschriebenen Definition einen Überblick über die jeweiligen Ausprägungen im Datensatz (z. B. 1 = lebend geboren; 2 = tot geboren).

... und Informationen zu den gesetzlichen Grundlagen

Im Bereich D werden dem Wissenschaftler – neben Vergleichen zwischen unterschiedlichen Erhebungen – die vollständigen, für die entsprechende Statistik relevanten Gesetzestexte zur Verfügung gestellt.

Praktische Erprobung des Metadatensystems für die zweite Jahreshälfte geplant

Wie die Forschungsdatenzentren der statistischen Ämter des Bundes und der Länder selbst ist auch die beschriebene Metadatenbank noch im Aufbau begriffen. Die ersten praktischen Erprobungen dieses Systems werden voraussichtlich in der zweiten Jahreshälfte beginnen können, so dass der Wissenschaft mit dieser Metadatenbank in absehbarer Zeit ein unverzichtbares Werkzeug zur Interpretation von Mikrodaten der amtlichen Statistik zur Verfügung stehen wird.

Die statistischen Ämter können somit nicht nur dem ständig wachsenden Informationsbedarf in unserer Gesellschaft Rechnung tragen, sondern sie bieten mit den Forschungsdatenzentren auch alle notwendigen Informationen für die wissenschaftliche Forschung aus einer Hand an, die Mikro- und Makrodaten ebenso wie die Daten über diese Daten.

Alexander Richter, Diplom-Demograph, und Dr. Stefan Weil sind Referenten im Referat Analysen und Prognosen, Forschungsdatenzentrum.