Deutsches Jugendinstitut

12.01.2021

**Stellungnahme des DJI zum Referentenentwurf „Gesetz zur Änderung des E-Government-Gesetzes und zur Einführung des Gesetzes für die Nutzung von Daten des öffentlichen Sektors“ vom 17.12.2020[1]**

Das Deutsche Jugendinstitut (DJI) begrüßt die Gesetzesinitiative zur Umsetzung von Strategien für Open-Data im Bereich von Forschungsdaten. Satzungsgemäß verfolgt das DJI die Aufgabe Daten und Fakten zum Aufwachsen von Kindern und Jugendlichen in Deutschland zu sammeln, zu dokumentieren und zu analysieren. Im Rahmen dieser Aufgabe werden vielfältige Datenerhebungen durchgeführt, darunter auch regelmäßig bundesweite Befragungsstudien bei diversen Personengruppen. Die dabei gesammelten Individualdaten werden im Institut aufbereitet und analysiert. Zu der Aufbereitung gehören z.B. die Prüfungen der Datenkonsistenz, aber auch die Anreicherung mit Kontextinformationen und mit Metainformationen. Die resultierenden Forschungsdaten liegen in sehr unterschiedlichen Formaten, z.B. Ton- und Bildaufnahmen von Alltagssituationen, Gesprächen und Diskussionsrunden, Transkriptionen eben dieser Aufnahmen, in standardisierte Formate konvertierte Befragungs- oder Beobachtungsergebnisse, Meta- und Prozessdaten von Datenerhebungen vor. Unsere Ausführungen im Folgenden nehmen jeweils nur diese Daten in den Blick[2].

In der Regel werden diese Befragungs- und Metadaten in maschinenlesbaren Formaten dann, über das beim Rat für Wirtschafts- und Sozialdaten (RatSWD) akkreditierte Forschungsdatenzentrum des DJI, für wissenschaftliche Zwecke weitergegeben. Die Rechtsgrundlage der Verarbeitung von personenbezogenen Daten ist regelmäßig die informierte Einwilligung der untersuchten Personen. Neben den Vorgaben der geltenden Datenschutzbestimmungen – insbesondere der Datenschutz-Grundverordnung, des Bundesdatenschutzgesetzes, der Landesdatenschutzgesetze sowie spezialgesetzlicher Regelungen – werden die forschungsethischen Regeln der Geistes- und Sozialwissenschaften und verwandter Disziplinen berücksichtigt. Forschungsdaten werden anonymisiert, sobald dies nach dem jeweiligen Forschungszweck möglich ist. Dazu werden besondere Merkmale entfernt oder vergröbert, so dass eine Identifikation einzelner Personen anhand der Daten nicht mehr oder nur mit unverhältnismäßigem Aufwand an Zeit, Kosten und Arbeitskraft möglich ist. Die folgende Stellungnahme geht davon aus, dass die geltenden Datenschutzbestimmungen vorrangig zu wahren sind (vgl. auch § 2 Abs. 4 DNG).

Das Institut folgt mit der Datenweitergabe im Grunde einem Open-Data-Ansatz und richtet sich dabei nach den FAIR-Prinzipien zur wissenschaftlichen Datennutzung. FAIR steht dabei

[1] Diese Stellungnahme wurde hauptsächlich von PD Dr. Susanne Kuger, Holger Quellenberg und Peter Furthmüller erstellt.

[2] Darüber hinaus werden im DJI zudem Daten genutzt zur eigenen Administration und zur Erfüllung unserer Pflichten als Arbeitgeber und Drittmittelempfänger. Diese Daten sind durch die regelmäßigen Berichterstattungen des Hauses wie dem DJI-Jahresbericht aggregiert öffentlich zugänglich. Die folgende Stellungnahme geht davon aus, dass diese Daten nicht unter „Forschungsdaten“ zu fassen sind.

als Akronym für: Findable, Accessible, Interoperable, Reusable – Prinzipien, die sich auch in dem vorliegenden Referentenentwurf zum EGovG und DNG finden. Die für die wissenschaftliche Nachnutzung freigegebenen Forschungsdaten können nach Anmeldung am Forschungsdatenzentrum des DJI und Zusicherung eines wissenschaftlichen Verwendungszwecks schon jetzt frei heruntergeladen werden. Auch von daher begrüßen wir eine weitere rechtliche Fundierung unseres Vorgehens durch den vorliegenden Entwurf.

Vor dem beschriebenen Hintergrund haben wir einige Anmerkungen zu dem Entwurf. Wir orientieren uns dabei an der Gliederung aus „B. Besondere Teil".

## Zu Artikel 1 (Änderung des EGovG)

### Zu Nummer 3, Buchstabe b:

Die Vorgaben zu maschinenlesbaren Datenformaten sind wahrscheinlich schwierig umzusetzen. In den empirischen Sozial- und Wirtschaftswissenschaften haben sich bestimmte Softwarepakete durchgesetzt, mit jeweils proprietären Dateiformaten. Diese Dateiformate haben in der Regel den Vorteil, neben den Rohdaten auch die relevanten Metadaten auf Ebene von Einzelvariablen ablegen zu können. Eine Trennung der Informationen in Einzeldateien wäre natürlich technisch möglich, jedoch würde es den Umgang und die Arbeit mit den Daten erschweren und im wissenschaftlichen Bereich eher auf Unverständnis stoßen. Die Frage nach den Datensatzformaten sollte deshalb eine Öffnung hin zu wissenschaftlich etablierten Standards erlauben.

Die Formulierung „unbearbeitete Daten" unterliegt weitreichender Interpretierbarkeit. Eine übliche Erstbearbeitung, wie sie oben beschrieben wird, enthält in der Regel reine Fehlerkorrekturen, Konsistenz- und Plausibilitätsprüfungen, so dass die Daten danach erstmalig auswert- und interpretierbar sind. Häufig ist auch eine Erstbearbeitung notwendig, um Anonymisierungen durchzuführen, ohne die die Daten nicht datenschutzkonform veröffentlicht werden dürfen (vgl. auch Punkt 3 c des Referentenentwurfs). Eine Bereitstellung vor dieser Erstbearbeitung kann daher zu weitreichenden Schwierigkeiten oder Widersprüchen führen und wäre ein Rückschritt hinter derzeitige Standards der Datenbereitstellung in OpenData Repositorien und (akkreditierten) Forschungsdatenzentren. Entsprechende Probleme zeigen sich vor allem in Fällen, in denen sich noch keine Standards etabliert haben, insbesondere bei innovativen Datenerfassungsmethoden (z.B. webscraping, log-/Prozess-Daten oder einigen Formen qualitativer Daten). Eine Formulierung, die sich darauf beruft, dass die Daten nach den gängigen Standards maximal unbearbeitet sein sollten, wäre zu bevorzugen.

### Zu Nummer 3, Buchstabe c:

Wir begrüßen ausdrücklich die Förderung der Entwicklung hin zu einer Kultur der „Open Science". Auch die Möglichkeit, in einem weiteren Portal Roh- bzw. Metadaten anzubieten begrüßen wir.

Eine verpflichtende Teilnahme an dem nationalen Metadatenportal GovData erscheint uns bisweilen jedoch kontraproduktiv. International gibt es inzwischen eine Vielzahl von fachlich

einschlägigen Katalogen und Verzeichnissen von Forschungsdaten. Diese sind – zumindest im wissenschaftlichen Bereich – in der Regel bekannt und werden innerhalb der Fachdisziplinen rege genutzt. Häufig ergibt sich die Spezialisierung zwingend aus den disziplinspezifischen Eigenheiten der jeweiligen Daten. Dies führt zu unterschiedlichsten Metadatenkatalogen und -beschreibungen, die jeweils Verwendung finden. So wird in der Geologie ein arktischer Bohrkern andere Merkmale (und somit Metadaten) aufweisen, als ein Länderdatensatz in den Wirtschaftswissenschaften; ein schulischer Leistungstest hat andere Eigenschaften als eine Elternbefragung oder eine CT-Aufnahme der menschlichen Lunge. Werden solche disparaten Datenbestände einem Metadatenschema zugeordnet, so dürfte der praktische Nutzen stark verkürzt sein.

Deshalb würden wir auch hier eine Öffnung hin zu verbreiteten Standards als Alternative zu der verpflichtenden Nutzung eines verbindlichen Metadatenportals begrüßen. Auch eine verpflichtende Beteiligung an der NFDI Initiative (Nationale Forschungsdateninfrastruktur), einer der darin erfolgreich geförderten Vorhaben und eventuellen Folgeinitiativen wäre eine Alternative.

In diesem Absatz zeigt sich die oben schon beschriebene Unschärfe des Begriffs der Rohdaten oder der „unbearbeiteten Daten“. Es bleibt unklar, was vom Gesetzgeber darunter verstanden wird.

Für die Sozialwissenschaften kann, wie oben beschrieben, festgehalten werden, dass schon in der Verwendung des Begriffs der „Rohdaten“ mit Punkt 3 c, dem Hinweis, dass die Daten anonymisiert werden sollten, in Konflikt steht. Weder in den vorranging geltenden Datenschutzbestimmungen noch im vorliegenden Gesetzesentwurf wird der Begriff der Anonymisierung auf eine Weise definiert, der Rechtsunsicherheiten für die Anbieter von sozialwissenschaftlichen Forschungsdaten ausräumen kann. Je nach Studiendesign, verfügbarem Zusatzwissen und technischer Entwicklung kann bei quantitativen und insbesondere bei qualitativen Individualdaten ein mehr oder weniger großes De-Anonymisierungsrisiko bestehen. Der Gesetzesentwurf scheint eine „absolute Anonymisierung“ zu fordern, § 3 Nr. 13, wonach eine De-Anonymisierung gänzlich ausgeschlossen sein muss. Der Bundesbeauftragte für Datenschutz und Informationsfreiheit (BfDI) hat kürzlich in einem Positionspapier festgestellt: „Eine absolute Anonymisierung derart, dass die Wiederherstellung des Personenbezugs für niemanden möglich ist, dürfte häufig nicht möglich sein und ist im Regelfall auch nicht gefordert. Ausreichend ist in der Regel, dass der Personenbezug derart aufgehoben wird, dass eine Re-Identifizierung praktisch nicht durchführbar ist, weil der Personenbezug mit einem unverhältnismäßigen Aufwand an Zeit, Kosten und Arbeitskraft wiederhergestellt werden kann.“ (BfDI 2020, S. 4)[3] Es wäre wünschenswert, wenn der Gesetzesentwurf diese Problematik aufgreifen und zum Begriff der Anonymisierung stärkere Rechtssicherheit schaffen würde.

Rohdaten liegen nur in seltenen Fällen absolut oder faktisch anonymisiert vor. Eine Anonymisierung ist bei Personenbefragungen zwingend mit einer Veränderung der Rohdaten verbunden. Bei einer Fragebogenerhebung entsprechen z.B. ausgefüllte Bögen oder ein Scan davon den Rohdaten, schon ein Übertragen der Information in eine rechteckige Datenmatrix

[3] Der Bundesbeauftragte für den Datenschutz und die Informationsfreiheit (2020): Positionspapier zur Anonymisierung unter der DSGVO unter besonderer Berücksichtigung der TK-Branche.

(Fälle in Zeilen, z.B. Schüler, und Variablen in Spalten, z.B. Schulnoten im Sport und Körpergröße), also z.B. das Festhalten der Ausprägung „180“ auf die Frage nach „Körpergröße in cm“, stellt eine Bearbeitung der Rohdaten dar.

Und hier wird dann auch eine zweite Unschärfe deutlich: Die Trennung von Individualdaten („Personen“) und Aggregatdaten („Tabellen“) erfolgt in dem Entwurf in der Regel nicht.

Im Metadatensystem GovData sind, nach unseren Recherchen, bisher überwiegend Tabellendaten, also aggregierte, zusammengefasste, abgeleitete, ausgewertete oder berechnete Daten abgelegt. Wissenschaftlich deutlich interessanter sind häufig aber die zugrundeliegenden Individualdaten, also die Daten, deren Auswertungen zu den Tabellen führen (z.B. Größe aller Kinder in einer Klasse). Gleichzeitig enthalten diese jedoch immer auch sensible Informationen, die eine Identifizierung bestimmter Befragter erlauben und deshalb nicht weitergegeben werden können. Bei diesen Individualdaten handelt es sich um die (unbearbeiteten) Rohdaten, die durch Bearbeitung zu anonymisierten Daten werden. Zu analysierbaren Forschungsdaten werden sie jedoch erst, wenn sie mit Metadaten verknüpft wurden.

Eine klarere Begriffsdefinition der verwendeten Begriffe sowie eine Festlegung auf die präferiert zu veröffentlichenden Daten (Rohdaten, ggf. bereinigte Individualdaten, oder aggregierte und aufbereitete oder ausgewertete Daten) sind unbedingt erforderlich.

**Zu Nummer 3, Buchstabe f:**

Eine Formulierung, die auf größtmögliche Aktualität bei der Veröffentlichung hinweist, würde die Bedeutung dieses Datenmerkmals für viele Nachnutzungsbereiche unterstreichen (Vgl. auch die in Absatz 4 benannten Fristen, die in § 19 genannt werden sollen: ähnliche Fristen sind auch für die nachfolgend laufenden Veröffentlichungen denkbar.) Diese Veröffentlichungsfristen sollten jedoch nur als Richtlinien gelten und Ausnahmen zulassen, wie sie schon jetzt in Zuwendungsnebenbestimmungen verschiedener Förderer üblich sind (etwa zum Zweck des Abschlusses von Qualifikationsarbeiten).

**Zu Artikel 2 (Neues DNG)**

**Zu § 2, Absätze 5 und 6**

Gerade die Begründung für die Verfügbarkeit von Daten in Bezug auf § 3 Absatz 3 wirft die Frage auf, warum gerade Schulen (Sekundarstufe abwärts) von der Veröffentlichung der Daten ausgenommen sind. Im Rahmen der Schulpflicht erfüllt das Schulsystem wichtige öffentliche Aufgaben, von deren transparenterer Darstellung und Dokumentation Forschung, Bildungsadministration und -praxis enorm profitieren würden. Gerade in Phasen hohen Zeitdrucks (z.B. Pandemiegeschehen, Probleme bei zentralen Leistungsprüfungen) wäre ein einfacherer und unmittelbarer Zugriff auf Rahmenbedingungen des Schulwesens und -alltags hilfreich.

**Zu § 3, Absatz 9**

Das Kennzeichen der Dynamik ist stark disziplin- und datenabhängig. Die Begrifflichkeit „häufig“ bedarf daher näherer Interpretation, da sie disziplinär sehr unterschiedlich ausgelegt ist: Über eine Wetterveränderung einmal pro Woche zu berichten wäre nicht häufig, über eine Arbeitslosenquote jedoch durchaus.

**Zu § 7, Absatz 2**

Wie oben zu Artikel 1 EGovG ausgeführt, verlieren Daten an Wert oder werden zumindest nur deutlich komplexer verarbeitbar, wenn auf gängige Formate zugunsten vollständig offener Formate verzichtet wird; dies gilt vor allem, wenn dadurch Daten von Metadaten getrennt gespeichert werden müssen. In Ergänzung zu vollständig offenen Formaten sollten daher auch dem jeweils aktuellen disziplinären Standard entsprechende Formate zulässig sein.

Aus den oben beschriebenen Gründen sollte auf die Vorgabe offener Datenformate verzichtet werden. Dafür könnten, entsprechend dem Stand der Technik, etablierte Formate favorisiert werden.

**Zu § 7, Absatz 4**

Aus den oben genannten Gründen (s. Argument zu 3c EGovG) sollte auf eine Verpflichtung zur Ablage von Metadaten ausschließlich in GovData verzichtet werden.