

Publikation von Forschungsdaten

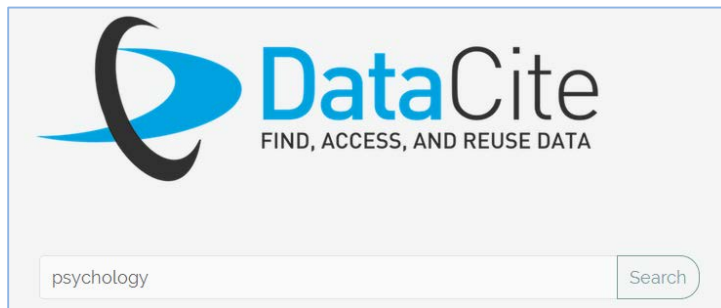
Vorstellung der Repositorienlandschaft in der Psychologie

I. Dehnhard, M. Kerwer, A. Günther, E. Weichselgartner
Leibniz-Zentrum für Psychologische Information und Dokumentation

Suche nach Repositorien

„DataCite is a leading global non-profit organisation that provides persistent identifiers (DOIs) for research data.“

Search DataCite registry to find datasets, software, images, and other research material.



Suchanfrage: „psychology“

Resource types: „Dataset“

Resource Types	
<input type="checkbox"/> Text	7,387
<input checked="" type="checkbox"/> Dataset	1,881
<input type="checkbox"/> Collection	1,012
<input type="checkbox"/> Image	627
<input type="checkbox"/> Other	525
<input type="checkbox"/> Audiovisual	233
<input type="checkbox"/> Software	78
<input type="checkbox"/> Film	40

Auswahl von Repositorien

Data Centers

<input type="checkbox"/> figshare Academic Research System	821
<input type="checkbox"/> DANS-EASY	545
<input type="checkbox"/> GESIS.ZPID	53
<input type="checkbox"/> Monash University	53
<input type="checkbox"/> ICPSR	52
<input type="checkbox"/> University of York	40
<input type="checkbox"/> Romania ADL Association	37
<input type="checkbox"/> Dryad	33
<input type="checkbox"/> GESIS Leibniz Institute for the Social Sciences	25
<input type="checkbox"/> Harvard Dataverse	24
<input type="checkbox"/> Forschungsdatenzentrum Deutsches Zentrum für Altersfragen (FDZ-DZA)	19
<input type="checkbox"/> Università del Salento	17
<input type="checkbox"/> ZENODO - Research. Shared.	17
<input type="checkbox"/> UCSD	16
<input type="checkbox"/> The State Archives	12

➤ Insgesamt **15** Datenrepositorien

Auswahl von Repositorien

Data Centers

<input type="checkbox"/> figshare Academic Research System	821
<input type="checkbox"/> DANS-EASY	545
<input type="checkbox"/> GESIS.ZPID =FDZ PsychData	53
<input type="checkbox"/> Monash University	53
<input type="checkbox"/> ICPSR	52
<input type="checkbox"/> University of York	40
<input type="checkbox"/> Romania ADL Association	37
<input type="checkbox"/> Dryad	33
<input type="checkbox"/> GESIS Leibniz Institute for the Social Sciences	25
<input type="checkbox"/> Harvard Dataverse	24
<input type="checkbox"/> Forschungsdatenzentrum Deutsches Zentrum für Altersfragen (FDZ-DZA)	19
<input type="checkbox"/> Università del Salento	17
<input type="checkbox"/> ZENODO - Research. Shared.	17
<input type="checkbox"/> UCSD	16
<input type="checkbox"/> The State Archives	12

- Insgesamt **15** Datenrepositorien
- Davon **7**, bei denen keine Datenübergabe von „Externen“ möglich
- Auswahl von **8** Datenrepositorien + OSF (Open Science Framework)

3. Art der Speicherung

1. **Unabhängigkeit & wissenschaftliche Professionalität**
2. **Persistenz (Dauer der Aufbewahrung)**
3. **Zugänglichkeit**
 1. **Öffentlich**
 2. **Kostenlos**
 3. **Einschränkbar**
4. **Nutzung persistenter Identifikatoren (PIDs)**
(Identifizierbarkeit)
5. **Klärung der Rechte**
6. **Nicht-öffentliche Speicherung möglich**

3. Art der Speicherung

Die Primärdaten sollen in digitaler Form¹¹ auf einem vertrauenswürdigen Repositorium bereitgestellt werden. Wichtige Qualitätsmerkmale eines vertrauenswürdigen Repositoriums sind¹²:

- Die wirtschaftliche bzw. ideologische **Unabhängigkeit** und wissenschaftliche **Professionalität** der bereitstellenden Institution;
- Die **Persistenz** der Daten: Es muss gewährleistet sein, dass die Daten über einen langen Zeitraum (mindestens 10 Jahre, im Idealfall jedoch deutlich länger) gesichert sind, und es muss geklärt sein, was im Falle einer Auflösung des Repositoriums mit den Daten geschieht;
- Die **Zugänglichkeit** der Daten: Es muss möglich sein, die Daten öffentlich und kostenlos abzurufen; es muss aber auch möglich sein, Zugangsbeschränkungen (im Sinne von „Scientific Use Files“) zu definieren (zur Diskussion einer eventuellen Zugangsbeschränkung siehe Abschn. 5);
- Die **Identifizierbarkeit** der Daten: Es muss ein persistenter Identifikator (z.B. eine persistente URL oder, falls möglich, ein DOI) vergeben werden;
- Die **Klärung der Rechte** an den Daten: Mit dem Speichern der Daten darf kein Abtreten der ausschließlichen Nutzungsrechte an Dritte verbunden sein (das einfache Nutzungsrecht, d.h. das Recht zur Archivierung und Vervielfältigung muss dem Betreiber des Repositoriums übertragen werden, damit der Betrieb des Repositoriums überhaupt möglich ist);
- Die Möglichkeit, Dateien bzw. Datensätze sowohl öffentlich als auch nicht-öffentlich abzuspeichern.

Schönbrodt, Gollwitzer, & Abele-Brehm, 2016

Anhang A

1. **Dauer der Aufbewahrung (Persistenz)**
2. **Verwendung nicht-proprietärer Datenformate**
3. **Dokumentation auf Variablenebene (Inhaltliche Verständlichkeit)**
4. **Nutzung PIDs(Auffindbarkeit)**
5. **Dokumentation**
 1. auf **Studienebene**
 2. auf **Datenobjektebene**
6. **Bereitstellung von**
 1. **Instrumenten**
 2. **Software**
 3. **Materialien**

1. Dauer der Aufbewahrung: Die DFG schreibt vor, die Primärdaten *mindestens* 10 Jahre aufzubewahren. Diese Regelung leitet sich aus der Guten Wissenschaftlichen Praxis ab, die insbesondere auf die Vermeidung von wissenschaftlichem Fehlverhalten abzielt und sicherstellen soll, dass in einem Verdachtsfall die Forschungsergebnisse nachvollzogen und geprüft werden können. Das impliziert nicht, dass die Daten nach diesem Zeitraum gelöscht werden *sollten*. Da es, außer bei Studien mit extrem großen Datenmengen, kein Argument für die Löschung anonymisierter Daten gibt, befürwortet die DGPs im Normalfall eine unbegrenzte Sicherung ohne zeitgebundenen "Löschmechanismus" oder „Löschzwang“.
2. Die Speicherung in einem nicht-proprietären Datenformat (wie z.B. csv-Dateien oder Text-Dateien) ist einem proprietären Dateiformat (.mat-Datei, SPSS-Datei, SAS-Datei) vorzuziehen damit die Transparenz nicht auf Besitzer(innen) ggf. teurer Spezialsoftware beschränkt wird. Entsprechend sollen die Daten so aufbereitet sein, dass sie ohne Spezialsoftware auslesbar und nutzbar sind. Zum Beispiel würde ein binärer Output aus einer unbekannten Experimentalsoftware dem Ziel der Forschungstransparenz nicht genügen; hier sind die Daten soweit aufzubereiten, dass sie in einem nachvollziehbaren Datenformat vorliegen. Wenn das nicht möglich ist, sollte klar dokumentiert werden, mit welcher Software die Daten auslesbar sind.
3. Neben der technischen Zugänglichkeit muss auch die inhaltliche Verständlichkeit der Daten gewährleistet sein. Alle Variablen müssen in einem digitalen Codebuch dokumentiert sein²⁶. Es muss klar sein, welche manipulierte bzw. gemessene Variable in der Publikation zu welcher Variablen im Datensatz gehört. Idealerweise liegen den Daten auch Analyseskripte bei (z.B. R-Skripte oder SPSS-Syntax), die die publizierten Ergebnisse reproduzieren.
4. Der Speicherort sollte einen persistenten Identifikator erhalten (z.B. eine persistente URL oder falls möglich, einen DOI). Das erlaubt eine einheitliche Zitation der Daten. Der Speicherort soll in der Publikation, die auf diesen Daten beruht, genannt werden, so dass Forschungsdaten auch gefunden werden können.
5. Neben der Dokumentation auf Variablenebene (Codebuch) ist auch eine Dokumentation auf Studienebene und auf Datenobjektebene (Dateien, Versionen, etc.) erforderlich. Das kann zum Beispiel durch eine README-Datei im Repository gelöst werden, die einen Überblick über die archivierten Dateien gibt, und gegebenenfalls Hinweise zur Reproduktion gibt (z.B., Welche Software ist notwendig? In welcher Reihenfolge sind die Skripte auszuführen? Wie soll der Datensatz zitiert werden?).
6. Es wird empfohlen, neben den Daten auch Instrumente, Software und Materialien

Schönbrodt, Gollwitzer, & Abele-Brehm, 2016

Anhang A

1. Dauer der Aufbewahrung (Persistenz)
2. Verwendung nicht-proprietärer Datenformate
3. Dokumentation auf Variablenebene (Inhaltliche Verständlichkeit)
4. Nutzung persistenter Identifikatoren (Auffindbarkeit)
5. Dokumentation
 1. auf Studienebene
 2. auf Datenobjektebene
6. Bereitstellung von
 1. Instrumenten
 2. Software
 3. Materialien

1. Dauer der Aufbewahrung: Die DFG schreibt vor, die Primärdaten *mindestens* 10 Jahre aufzubewahren. Diese Regelung leitet sich aus der Guten Wissenschaftlichen Praxis ab, die insbesondere auf die Vermeidung von wissenschaftlichem Fehlverhalten abzielt und sicherstellen soll, dass in einem Verdachtsfall die Forschungsergebnisse nachvollzogen und geprüft werden können. Das impliziert nicht, dass die Daten nach diesem Zeitraum gelöscht werden *sollten*. Da es, außer bei Studien mit extrem großen Datenmengen, kein Argument für die Löschung anonymisierter Daten gibt, befürwortet die DGPs im Normalfall eine unbegrenzte Sicherung ohne zeitgebundenen "Löschmechanismus" oder „Löschzwang“.
2. Die Speicherung in einem nicht-proprietären Datenformat (wie z.B. csv-Dateien oder Text-Dateien) ist einem proprietären Dateiformat (.mat-Datei, SPSS-Datei, SAS-Datei) vorzuziehen damit die Transparenz nicht auf Besitzer(innen) ggf. teurer Spezialsoftware beschränkt wird. Entsprechend sollen die Daten so aufbereitet sein, dass sie ohne Spezialsoftware auslesbar und nutzbar sind. Zum Beispiel würde ein binärer Output aus einer unbekannten Experimentalsoftware dem Ziel der Forschungstransparenz nicht genügen; hier sind die Daten soweit aufzubereiten, dass sie in einem nachvollziehbaren Datenformat vorliegen. Wenn das nicht möglich ist, sollte klar dokumentiert werden, mit welcher Software die Daten auslesbar sind.
3. Neben der technischen Zugänglichkeit muss auch die inhaltliche Verständlichkeit der Daten gewährleistet sein. Alle Variablen müssen in einem digitalen Codebuch dokumentiert sein²⁶. Es muss klar sein, welche manipulierte bzw. gemessene Variable in der Publikation zu welcher Variablen im Datensatz gehört. Idealerweise liegen den Daten auch Analyseskripte bei (z.B. R-Skripte oder SPSS-Syntax), die die publizierten Ergebnisse reproduzieren.
4. Der Speicherort sollte einen persistenten Identifikator erhalten (z.B. eine persistente URL oder falls möglich, einen DOI). Das erlaubt eine einheitliche Zitation der Daten. Der Speicherort soll in der Publikation, die auf diesen Daten beruht, genannt werden, so dass Forschungsdaten auch gefunden werden können.
5. Neben der Dokumentation auf Variablenebene (Codebuch) ist auch eine Dokumentation auf Studienebene und auf Datenobjektebene (Dateien, Versionen, etc.) erforderlich. Das kann zum Beispiel durch eine README-Datei im Repository gelöst werden, die einen Überblick über die archivierten Dateien gibt, und gegebenenfalls Hinweise zur Reproduktion gibt (z.B., Welche Software ist notwendig? In welcher Reihenfolge sind die Skripte auszuführen? Wie soll der Datensatz zitiert werden?).
6. Es wird empfohlen, neben den Daten auch Instrumente, Software und Materialien

Schönbrodt, Gollwitzer, & Abele-Brehm, 2016

Datenrepositorien - Überblick

Datenrepositorium	Betreiber	Sitz (Jurisdiktion)
figshare	Macmillan Publishers	United Kingdom
DANS EASY	Dutch Academy KNAW & funding organization NOW	Niederlande
PsychData	Leibniz-Institut für Psychologische Information & Dokumentation (ZPID)	Deutschland
ICPSR	Inter-University Consortium of Political and Social Science	USA
Dryad	Dryad Digital Repository	USA
GESIS	Leibniz-Institut für Sozialwissenschaften	Deutschland
Harvard Dataverse	Institute for Quantitative Social Science (IQSS)	USA
Zenodo	CERN (Europäische Organisation für Kernforschung)	Schweiz
Open Science Framework	COS (Center for Open Science)	USA

Datenrepositorien - Überblick

Datenrepositorium	Betreiber	Sitz (Jurisdiktion)
figshare	Macmillan Publishers	United Kingdom
DANS EASY	Dutch Academy KNAW & funding organization NOW	Niederlande
PsychData	Leibniz-Institut für Psychologische Information & Dokumentation (ZPID)	Deutschland
ICPSR ➡ openICPSR	Inter-University Consortium of Political and Social Science	USA
Dryad	Dryad Digital Repository	USA
GESIS	Leibniz-Institut für Sozialwissenschaften	Deutschland
Harvard Dataverse	Institute for Quantitative Social Science (IQSS)	USA
Zenodo	CERN (Europäische Organisation für Kernforschung)	Schweiz
Open Science Framework	COS (Center for Open Science)	USA

Datenrepositorien - Überblick

Datenrepositorium	Betreiber	Sitz (Jurisdiktion)
figshare	Macmillan Publishers	United Kingdom
DANS EASY	Dutch Academy KNAW & funding organization NOW	Niederlande
PsychData	Leibniz-Institut für Psychologische Information & Dokumentation (ZPID)	Deutschland
ICPSR ➡ openICPSR	Inter-University Consortium of Political and Social Science	USA
Dryad	Dryad Digital Repository	USA
GESIS ➡ datorium	Leibniz-Institut für Sozialwissenschaften	Deutschland
Harvard Dataverse	Institute for Quantitative Social Science (IQSS)	USA
Zenodo	CERN (Europäische Organisation für Kernforschung)	Schweiz
Open Science Framework	COS (Center for Open Science)	USA

Alle der genannten Repositorien bieten:

- Persistenz (mind. 10jährige Aufbewahrung)
- Öffentliche Zugänglichkeit
- Nutzung persistenter Identifikatoren
- Kein Abtreten der ausschließlichen Nutzungsrechte
- Bereitstellung zusätzlicher Materialien

Repositorien - Vergleich

	figshare	DANS (EASY)	PsychData	openICPSR	Dryad	datorium	Harvard Dataverse	Zenodo	OSF
Kostenfreie Zugänglichkeit	✓	✓	✓	✓/✗	✓	✓	✓	✓	✓
Einschränkbare Zugänglichkeit	✗	✓	✓	✗/✓	✗	✓	✓	✓	✗
Nicht-öffentliche Speicherung	✓	✗	✗	✓	✗	✗	✓	✗	✓
Umwandlung in nicht- proprietäre Formate	✗	✗	✓	✗	✗	✗	✓	✗	✗
Variablendokumentation	✗	✗	✓	✗	✗	✗	✗	✗	✗
Studiendokumentation	✗	✓	✓	✓	✓	✓	✓	✓	✗
Dokumentation von Datenobjekten	✗	✓	✓	✓	✓	✓	✓	✗	✗

Repositorien - Vergleich

	figshare	DANS (EASY)	PsychData	openICPSR	Dryad	datorium	Harvard Dataverse	Zenodo	OSF
Kostenfreie Zugänglichkeit	✓	✓	✓	✓/✗	✓	✓	✓	✓	✓
Einschränkbare Zugänglichkeit	✗	✓	✓	✗/✓	✗	✓	✓	✓	✗
Nicht-öffentliche Speicherung	✓	✗	✗	✓	✗	✗	✓	✗	✓
Umwandlung in nicht- proprietäre Formate	✗	✗	✓	✗*	✗	✗*	✓	✗	✗
Variablendokumentation	✗	✗	✓	✗*	✗	✗*	✗	✗	✗
Studiendokumentation	✗	✓	✓	✓	✓	✓	✓	✓	✗
Dokumentation von Datenobjekten	✗	✓	✓	✓	✓	✓	✓	✗	✗

*Bei ICPSR und GESIS: ✓

Repositorien - Vergleich

	figshare	DANS (EASY)	PsychData	openICPSR	Dryad	datorium	Harvard Dataverse	Zenodo	OSF
Kostenfreie Zugänglichkeit	✓	✓	✓	✓/✗	✓	✓	✓	✓	✓
Einschränkbare Zugänglichkeit	✗	✓	✓	✗/✓	✗	✓	✓	✓	✗
Nicht-öffentliche Speicherung	✓	✗	✗	✓	✗	✗	✓	✗	✓
Umwandlung in nicht- proprietäre Formate	✗	✗	✓	✗*	✗	✗*	✓	✗	✗
Variablendokumentation	✗	✗	✓	✗*	✗	✗*	✗	✗	✗
Studiendokumentation	✗	✓	✓	✓	✓	✓	✓	✓	✗
Dokumentation von Datenobjekten	✗	✓	✓	✓	✓	✓	✓	✗	✗

*Bei ICPSR und GESIS: ✓

Fazit:

- Alle 9 Datenrepositorien erfüllen mind. 9 der 12 Kriterien
- Kein Datenrepositorium erfüllt alle Kriterien
- Unterschiede betreffen vor allem Nicht-öffentliches Speichern und Kriterien, die technische & inhaltliche Verfügbarkeit sicherstellen

Kritische Überlegungen :

- Wieviele der Dateien können in 2, 5 oder 10 Jahren noch geöffnet werden?
- Wieviele der Daten werden in 2, 5 oder 10 Jahren noch interpretierbar sein?
- Sind Datensätze auffindbar bei individuell gestalteten Dokumentationen?
- Kann die Verknüpfung von Artikel und Datensatz langfristig garantiert werden?
- Braucht es die Festlegung von Standards zur Dokumentation (auch für Variablen)?
- Braucht es Review Prozeduren für Daten?

Vielen Dank für Ihre Aufmerksamkeit!

www.psychdata.de

psychdata@zpid.de

Literatur:

Schönbrodt, F., Gollwitzer, M., & Abele-Brehm, A. (2017). Der Umgang mit Forschungsdaten im Fach Psychologie: Konkretisierung der DFG-Leitlinien. Psychologische Rundschau, 68, 20–35. doi:10.1026/0033-3042/a000341.