gesis
Leibniz-Institut für Sozialwissenschaften

Forschungsdatenmanagement in den Sozialwissenschaften

Uwe Jensen,
GESIS – Leibniz-Institut für Sozialwissenschaften

Workshop „Fachübergreifender Austausch zum Forschungsdatenmanagement“
DFG-Projekt LABIMI/F 25. Juni 2012, Berlin

Leibniz-Gemeinschaft

Sozialwissenschaften & Daten

- Sammelbegriff: Anthropologie > Volkswirtschaftslehre
- Gemeinsamkeit: Frage gesellschaftl. Zusammenlebens
  soziale Gerechtigkeit, Analyse politischer Systeme, ... Kernfächer: Soziologie / Politologie
- Forschungsdaten <> empirische Sozialforschung Kennzeichen:
  - Momentaufnahmen gesellschaftlicher Zustände
  - analytischer Zusatzwert durch systematische Einordnung in Zeitvergleiche
  - Sekundäranalyse (1960zig.) > international vergleichende Projekte
  - Forschungsinteresse <> Gründung von SOWI Datenarchiven (Roper 1957, ZA 1960, NL …)
  - Strukturen von Datenerhebungsprojekte:
    - Regel > kleinteilige Einzelerhebungen // langfristige Erhebungsprogramme (ALLBUS, SOEP, EVS, ISSP, SHARE....) > brauchen feste Strukturen & geregeltes FDM
    - Datenerhebung: kommerzielle Meinungs- und Marktforschungsfirmen (Standards)
  - Small data size – but large complexity > Einzelstudie > Zeitreihen > ... (N sample x N Erhebungen)

FDM in Sozialwissenschaften

Forschungsdatenmanagement ist abhängig von Nutzungsszenario der Daten
z. B. hinsichtlich:

- **Wer ist Nutzer**? Projekt intern > Breite wissenschaftliche Nutzung > Öffentlichkeit ...
- **Welche Daten?** Originaldaten > Subsets > eigene & fremde Daten > Analysedaten ...
- **Wofür?** Forschungsbericht > Re-Analyse > Replikation > Sekundäranalyse ...
- **Wo/Wie?** Ablegen > intern oder extern Sichern (10J) > Archivieren > Vertrieb (Rechte)

LABIMI/F: Anforderungen in zwei Anwendungsschwerpunkten des FDM:

- Archivierung zur Beweissicherung (> Gute wissenschaftliche Praxis)
- Archivierung zur Nachnutzung durch interne Datenproduzenten & externe Forscher

Entsprechende Aspekte des FDM in Sozialwissenschaften anhand 4 Phasen im Datenzyklus

1. Studien planen & Daten erheben
2. Daten aufbereiten & dokumentieren
3. Studien archivieren & Daten registrieren
4. Studien und Daten recherchieren

1. Studien planen & Daten erheben

FDM Aspekte (> Forschungsvorhaben > Datenmanagementplan)

- **Recherche & Exploration vorhandener Daten** – Warum Teil des FDM?
  - Nutzung vorhandener Daten > keine Kosten & Zeitaufwand für eigene Erhebung
  - Auf Forschungsstand aufbauen > Sekundäranalysen
  - Qualitätssicherung: Vorbereitung neuer Untersuchung („Aus alten Daten lernen“) …
- **Überlegungen zur Archivierung & Nachnutzung von Daten**
  - **Art & Inhalt Dokumentationen?** > Datenbeschreibung, Methodenbericht, Syntaxdateien aus Statistikprogramm,
  - **Deskriptive Metadaten?** Berichte, techn. Standards, fachliche Standards (Klassifikationen ...)
  - **Partner?** Wo archivieren? interne Repository <> externer Datenservice / Spezifische Domäne? Was, wie lange?
  - **Bereitstellung?** Zugangsregelung, Datenschutz, Urheber-, Vertriebsrechte, …
- **Verantwortlichkeiten im FDM**
- **Personal- und Sachkosten im FDM**
  - **Kostenschätzung**: Je nach Struktur, Größe, Organisationsform <> eingeführte Praxis: FDM Prozesse & Regeln
  - **Qualitätsgesicherte Aufbereitung & Kontrolle** aller Materialien (Daten, Studienmaterial) für die Nachnutzung
  - **Datenorganisation & -sicherung > Digitalisierung von Material > Datenschutz & Anonymisierung ...**

1. Studien planen & Daten erheben

Methodische Formen & Quellen:

- standardisierte Umfragen, qualitative Interviews, amtliche Statistik, Prozessdaten …

FDM Anforderungen:

- **Datenschutz** bei der Erhebung & Verarbeitung personenbezogener Daten
  > komme darauf zurück
- **Datenerhebung & Dokumentation der Feldarbeit**
  - **Messinstrument**:
    Entwicklung & Pre-Test (Fragebogen, Leitfaden, ... – Interviewerschulung)
  - **Stichprobe**:
    methodische Verfahren zur Auswahl Befragter
  - **Erhebungskontext**:
    Metadaten (Paradaten) über Erhebungssituation, den Beteiligten, das Messinstrument , ...
    > Mittel der Qualitätskontrolle und -entwicklung (z. B. Share)
  - **Basis:**
    Qualitätsstandards der Marktforschung (ISO 20252:2006); ethische Prinzipien (ESOMAR Kodex)

2. Daten aufbereiten & dokumentieren

FDM Aspekte, um Daten projektunabhängig zu nutzen

- **Transparente, nachvollziehbare Dokumentation der Daten & Entstehungskontext**
  z. B. quantitative Daten aus Umfragen:
  1. F.-Projekt > **Studienkonzept**, **Methodendesign, Fragebogenentwicklung**,
  2. F.-Projekt > **Datendefinition** (Datenstruktur; Variablen) <> **Fragebogen** (Originalsprache > Übersetzungen),
  3. Erhebungsinstitut > **Feldbericht / Methodenbericht** … **Datenmodifikationen in Datenaufbereitung** > Datenkontrollen & Fehlerbereinigung > Konstruktion von Variablen, ...
  4. F.-Projekt > **Datenanalysen** > Syntaxdateien stat. Programme > Forschungsbericht , …

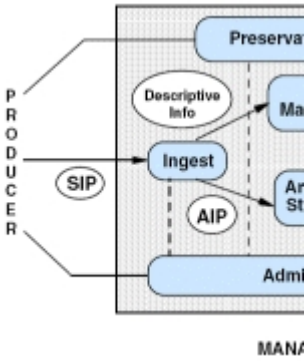
- **Einsatz von Metadatenstandards, Klassifikationen & Normen in der Dokumentation**
  - **Standards zur Verbesserung der Vergleichbarkeit von Daten**
    - Standarddemographie / Klassifikationen: Berufe, Ausbildungssystem ... / Skalen, Indikatoren, ...
    - ISO Normen für geografische Einheiten (ISO 3166), Sprachen (ISO 639-1:2002) ...
  - **Metadatenstandards > Nutzung in Großprojekten <u. o.> Dateneinrichtungen (FDZ, Datenarchiv)**
    - DDI Codebuchversion (v2.5) / DDI Life-Cycle (v3.1) > Daten der Empirischen Sozialforschung
    - SDMX (Statistical Data & Metadata Exchange) > Daten amtlichen Statistik (komplementäre Nutzung von DDI / SDMX)

3.1 Studien archivieren & Daten registrieren

GESIS Datenarchiv für Sozialwissenschaften (DAS) - Dienste:

- **Vorbereitung der Archivierung mit Datenproduzenten**
  - **Akquisition & Beratung zum Datenmanagement**:
  - **Vertragliche Regelung der Archivierung** (Nutzungsrechte, Zugang …)
  - **Material Datengeber:** Dokumentierter Datensatz, Erhebungsinstrument, Methodenbericht
- **Aufnahme der Studie ins Archiv**

  Standardregeln & -prozesse <> Heterogenität der Materials
  - Studienbeschreibung: inhaltliche, methodische, technische Charakteristika
  - Versionierung u. Vergabe eines persistenten Identifikators (DOI)
  - Technische und formale Kontrollen
  - Kontrollen zum Datenschutz > ggf. korrigierende Aufbereitungen
  - Sicherung aller Originaldateien und Materialien



- **Langzeitarchivierung:**
  - **Sicherung aller Originaldateien und Materialien im Archivspeicher**
  - **Substanzerhaltung:** räumlich getrennte & redundante Datenhaltung, diverse Speichertechnik - regelmäßige Medienmigration
  - **Erhalt von Nutzbarkeit & Interpretierbarkeit:** Migrationsstrategien (Überbrückung > Emulation bzw. Virtualisierung)

3.2 Studien archivieren & Daten registrieren

- **DAS Service in der Standardarchivierung**
  - einfache Aufbereitung (Korrektur von Daten u. Metadaten) u. Bereitstellung
- **DAS Mehrwertdienste > ausgewählte komparative Studien**
  - **Datenaufbereitung**:
    - Standardisierung von Datenstrukturen, Harmonisierung von Variablen,
    - Integration / Kumulation von Einzeldatensätzen (zeit- und / oder ländervergleichend),
    - Ergänzung mit Kontextdaten / Aggregatdaten
  - **Dokumentation**: Umfassende Produktion strukturierter Metadaten
    - z.B. vollständige u. multilinguale Frage- und Antworttexte, Interviewerанweisung,
    - Anmerkungen zur Datenqualität auf Variablenebene;
    Metadatenstandard > aktuell DDI Codebuchversion > Migration Produktionslinie nach DDI Life-Cycle
  - **Produkte**: Aufbereitete Daten + Dokumentationen
    Variablenreports (integrierte Dokumentation: Fragen, Daten, Datenaufbereitung), Methodenberichte, Originalfragebögen
  - **Service u. Forschung:** i.d.R. durch GESIS FDZs

FDZ ALLBUS
FDZ Amtliche Mikrodaten
FDZ Internationale Umfrageprogramme
FDZ Wahlen

ALLBUS
EUROBAROMETER
GLES
CSES
ISSP
European Values Study

gesis Leibniz Institute for the Social Sciences

Show all studies | Eurobarometer | EVS | ISSP | Studies Eastern Europe | Election Studies Germany | Politbar

DESCRIPTION | TABULATION | ANALYSIS

- European Values Study (EVS)
  - EVS 1981-2008 Longitudinal Data File
  - EVS 2008 - 4th wave
    - EVS 2008: Integrated Dataset
      - Metadata
      - Variable Description
        - Archive and ID Variables
        - [ZA4800] Weight
        - Perceptions of Life
        - Politics and Society

ZA4800: EVS 2008: Integrated Dataset

ZA4800 Datafiles and Documentation (download via data catalogue)

Variable v300: environment: humans were meant to rule over nature (Q85F)

LITERAL QUESTION

Q85

<SHOW CARD 85 – READ OUT AND CODE ONE ANSWER PER LINE!>

...t, can you tell me whether you agree

European Values Study - ZA4775 - v300 - Mozilla Firefox

info1.gesis.org/evs/variables/qdb.asp?db=QEVS2008&i...

Display this Variable in Extended Variable Overview

Dataset ZA4775: Croatia - EVS 2008

Variable v300: environment: humans were meant to rule

Q85

<POKAŽI KARTICU 85 - ČITAJ REDOM I IZABERI SAM

Pročitat ću Vam neke tvrdnje o okolišu. Možete li mi reći

uopće ne slažete?

Q85.F Ljudima je suđeno vladati ostatkom prirode.

1 u potpunosti se slažem
2 slažem se
3 ne slažem se
4 uopće se ne slažem
8 ne znam
9 nema odgovora

Note:

Missing values from field questionnaire recoded into -1 to

KARTICA 85

European Values Study - ZA4754 - v300 - Mozilla Firefox

info1.gesis.org/evs/variables/qdb.asp?db=QEVS2008&id=ZA4754&var=v300&lang=org

Display this Variable in Extended Variable Overview

Dataset ZA4754: Austria - EVS 2008

Variable v300: environment: humans were meant to rule over nature (Q85F) environment: humans were meant to rule over nature (Q85F

Q85

<Bitte Liste 85 vorlegen – VORLESEN UND PRO ZEILE JEWEILS EINE ANTWORT ANKREUZEN!>

Ich werde Ihnen nun einige Aussagen zum Thema Umwelt vorlesen. Bitte sagen Sie mir jedesmal, ob Sie voll und ganz zustimmen, zusti

stark ablehnen.

Q85.F Die Menschen sind dazu bestimmt, über die Natur zu herrschen

1 stimme voll und ganz zu
2 stimme zu
3 lehne ab
4 lehne voll und ganz ab
8 weiß nicht
9 keine Antwort

Note:

Missing values from field questionnaire recoded into -1 to -5 values in IDS and NDS.

LISTE 85

You

3.3 Studien archivieren & Daten registrieren

Veröffentlichung & Zitation von Daten als eigenständige Publikationsform

- **Nutzung v. Persistent Identifier** (Speicherort unabhängig > dauerhafte Infrastruktur)
- **Verschiedene PID Systeme**: Handle (Geistesw., Psycholinguistik), URN (Bibliotheken), PURL, …
- DataCite nutzt **DOI** (**Digital Object Identifier**)
  Anwendung in Klima, Medizin, SozWiss., Pädagogik, Biodiversität (Vergabe: TIB, ZBMED, ZBW, GESIS)

da|ra Datenregistrierung für Sozial- & Wirtschaftswissenschaften

- **Betrieb:** GESIS & ZBW (Zentralbibliothek Wirtschaftswissenschaften)
- **Metadatenstandard** DataCite : > mandatory fields für alle Datenpublikationen
- **da|ra Metadatenschema** (> DDI) > optionale Felder zur Spezifikation domainspezifischer Metadaten

Gewinn:

- **langfristige & kompakte Zitation (> Impact) & Zugänglichkeit der Daten**
- **Anerkennung: Datenproduktion als eigenständige Forschungsleistung**
- **Sichtbarkeit von Forscher & Produkt**

4. Daten recherchieren & bereitstellen

Datenquellen & Datenanbieter:

- Stat. Ämter; Verwaltungen etc. (amtliche Statistik; Prozessdaten)
- internationale Organisationen (Statistiken, Indikatorensysteme)
- Forschungsdatenzentren > RatSWD (www.ratswd.org)
- Datenarchive -> GESIS Datenarchiv für Sozialwissenschaften
  > europaweit vernetzt > CESSDA (www.cessda.org)

Datenservice des GESIS Datenarchiv:

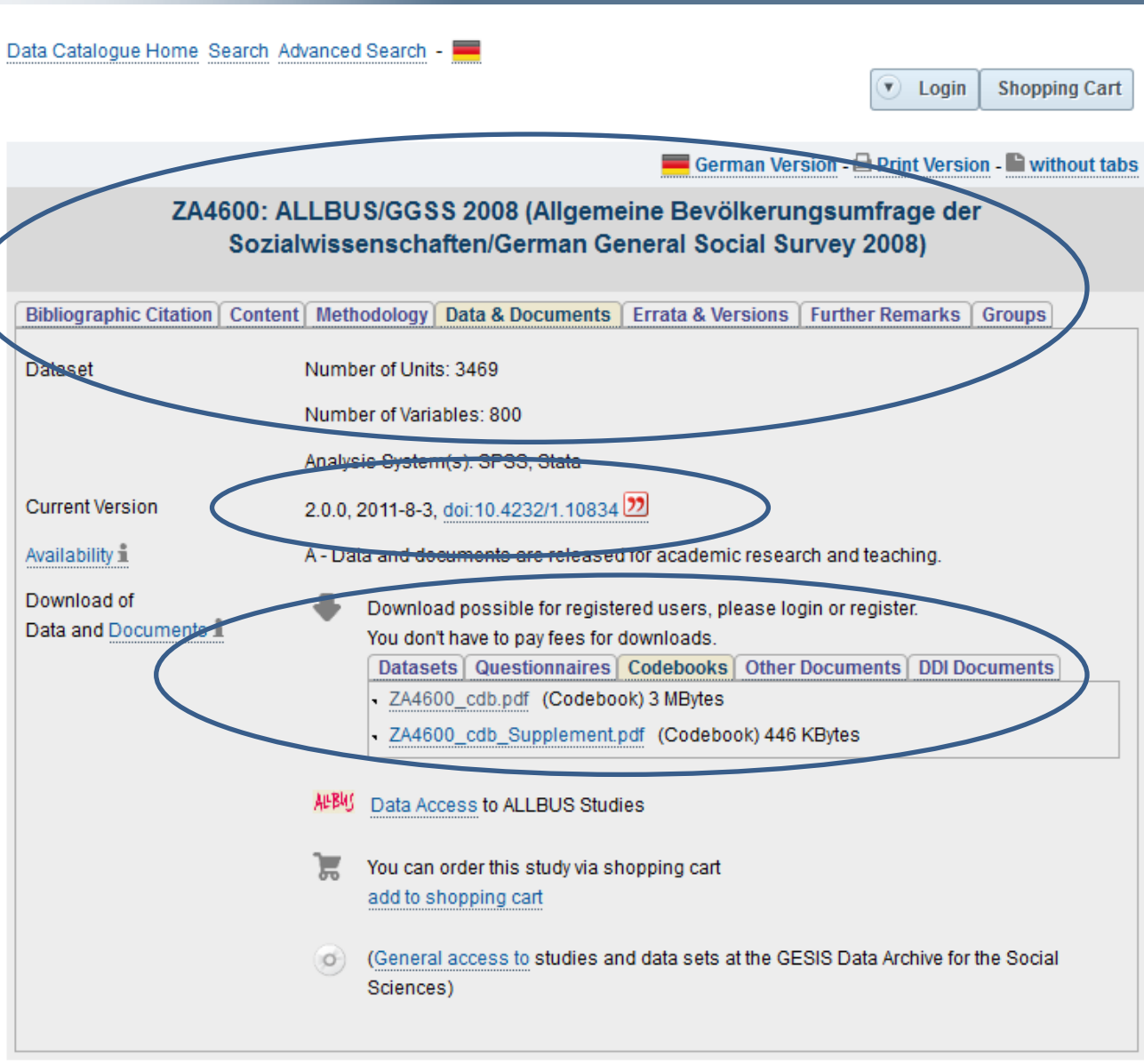
- Datenzugang über Online-Portale und (individuelle) Bereitstellung auf Datenträgern oder per ftp
  - **Datenbestandskatalog** (DBK) (Studienerschließung & Download (Daten, Dokumentationen)
  - **ZACAT**, **HISTAT** (Variablenrecherche, Online-Analyse, Download),
  - **sowiport** (Literatur & Daten)

Bsp.: **Erschließung & Präsentation von DDI basierten Metadaten einer Datenpublikation „doi:10.4232/1.10834“**

Oben: Kontext

Mitte:
Datensatz: Version & doi

Unten: Download
Daten & Dokumentation



F1. Welche Unterschiede in der Finanzierung von FDM?

Nutzungsszenario & Unterschiede? Archivierung zur Beweissicherung <> zur Nachnutzung projektintern + extern

Finanzierung FDM:

- I.d.R. keine explizite FDM Finanzierung in 3.Mittelprojekten > z. B. hinsichtl. Datendokumentation <> Nachnutzung / Archivierung > aber wachsendes Bewusstsein für Anforderungen & Kosten (z. B. KII-Gutachten)

A. Projektförderung der DFG - mit Aussagen zum Datenmanagement in 3 Bereichen

1. Allgemein für datenproduzierende Projekte > **Maßnahmen FDM zu Daten, die für Nachnutzung geeignet sind**

… um Daten nachhaltig zu sichern … ggf. für erneute Nutzung bereit zu stellen … existierende Standards
… Angebote von Datenrepositorien der Fachdisziplin nutzen (DFG 2012: 6)

2. Langfristvorhaben in den Geistes- und Sozialwissenschaften > FDM Panelstudien

… praktischen Maßnahmen zur Pflege eines Panels … zum kontinuierlichen Datenmanagement
… institutionelle Vorkehrungen um Weiterführung der Studien unter anderen Personen zu sichern.
Die Datendokumentation muss die langfristige Nutzung der Daten ermöglichen.“ (DFG 2011: 3)

3. DFG Sonderforschungsbereiche > zusätzliche Unterstützung FDM

- Service-Projekte im Bereich Informationsmanagement und Informationsinfrastruktur (INF)
- Konzeptrealisierung > Bereitstellung Infrastruktur > Langzeitarchivierung & kollaboratives Arbeiten

B. Institutionelle Förderung von Infrastruktureinrichtungen <> Schaffung von Nachhaltigkeit &

- **Datenarchiv der GESIS** - Leibniz-Institut für die Sozialwissenschaften > langfristige Bund-Länder-Finanzierung
- **Neue Forschungsdatenzentren** > überwiegend Anschubfinanzierung BMBF > in Mutter-Institute überführt (gleiche Finanzierung)
  z. B. FDZ SOEP am DIW Berlin / FDZ PsychData am ZPID Trier (Institute der Leibniz-Gemeinschaft))
- Forschungsdatenzentren der Länder & Statistisches Bundesamt müssen aber Kosten z.T. über Gebühren aus Datenservice decken
- **FDZ pairfam** (Basis DFG-Langfristvorhaben) Förderzeitraum: max. 12 Jahren

F2. Auswirkungen Datenschutzanforderung auf FDM?

Archivierung zur Beweissicherung <> Nachnutzung projektintern + extern

Empirische SozF. > Eingespielte Verfahren Teil der Qualitätssicherung von Erhebungsinstituten <> Forschungsprojekt <> Datenarchiv

- **Datenerhebung:** (Umfrage, Interviews)
  - **Stichprobenziehung** > Kontaktinformationen > frei zugänglich ? / >
  - **Einwilligungserklärung**; Aufklärung zu Verarbeitung & Nutzung für einen wissenschaftlichen Zweck (Archivierung eingeschlossen?)
- **Archivierung: Sind technische & administrative Voraussetzung zur Archivierung sensibler Daten gegeben?**
  - GESIS-Datenarchiv: **Archivierungsvertrag > Datenschutzprüfung > Nutzungsvertrag <> Zugangsklasse / techn. Infrastruktur**
  - Weitergabe von Projektdaten an ein Archiv muss in besonderen Fällen auch rechtlich geprüft werden (Eigentümer & Datenschutz)
- **Datenzugang (Nutzung durch Dritte): allgemein Bandbreite von … bis**
  - **völlig offener unbeschränkter** Zugang für jedermann (z.B. Public Use Files – absolut anonymisiert)
  - **eingeschränkte Nutzu**ng ausschließlich für wissenschaftliche Zwecke (z.B. Scientific Use Files – faktisch anonymisiert)
  - **hochrestriktive Nutzung**, z. B. geschützte Räumen in Dateneinrichtung (formal anonymisiert – Personendaten getrennt gehalten)
  - bzw. **auch vollkommener Ausschluss** der Nutzung durch Dritte

Je nach Studientyp und Datenquelle typische Anforderung > aber immer Einzelprüfung notwendig:

- **Einfache Querschnittsstudie (Einzelbefragung)**
  - Relativ geringer Aufwand
  - nach Datenaufbereitung (vollständig, faktisch) anonymisiert > dann keine Zugangsrestriktionen
  - Anonymisierungsmaßnahmen nötig: je größer Stichprobe, Anzahl demographischer Angaben bzw. Merkmalskombinationen
- **Wiederholungsbefragungen** (Panelstudien)
  - Größer Aufwand; besonderer Schutz der Kontaktdaten der Panelteilnehmer (Trennung von Daten);
  - Wenn nach Datenaufbereitung (vollständig, faktisch) anonymisiert > dann keine Zugangsrestriktionen
- **Qualitative Interviews**
  - erheblicher Aufwand bei der Anonymisierung ….
- **Daten statistischer Ämter und Verwaltungen** (Prozessdaten & Daten amtl. Statistik; Arbeitslose, Rentendaten, …)
  - erheblicher administrativer & technischer Infrastrukturaufwand;
  - Div. Nutzungsszenarien (On-Site – Safe-Center – kontrolliertes Fernrechen (formal anonymisierte Daten))
  - Transformationen der Daten, um extern für wissenschaftliche Zwecke zugänglich zu sein (Scientific Use Files)

F.3 Welche Aspekte wirken positiv auf die Akzeptanz der zusätzlichen Aufgaben durch ein FDM bei Forschern?

Archivierung zur Beweissicherung <> Nachnutzung projektintern + extern

- **Zitationsfähigkeit von Forschungsdaten** <> leichte Zugänglichkeit
- **Nachweismöglichkeit des Impacts** durch Datenzitationsraten
- **Anerkennung:** Datenproduktion &-dokumentation wertvolle Forscherleistung
- Weitere damit verbundene Faktoren:
  - **FDM unterstützte Datenqualität** erhöht **Reputation** von Projekten & Forschern,
    z. B. durch Transparenz & intersubjektiven Überprüfbarkeit der Daten & Forschungsergebnisse
    > Wettbewerbsvorteil im Ringen um Fördergelder (national / EU > SHARE, ESS)
  - **Kooperation** zwischen **Forschungsvorhaben und datenhaltender Einrichtung** zur Entwicklung & zum Austausch von projektspezifischem FDM Know-How
    > Projektspezifische Leitlinien zu Workflows, Metadatenhandling, Tools, Techniken, Verfahren ….
    > Arbeitsentlastung durch frühzeitige und strukturierte Abläufe und Regeln besonders bei mehreren Projektpartnern
  - Unterstützung von **Data Sharing & verbreiterte Datenbasis** für Sekundäranalysen
  - **Bewusste Beschreibung zum Umgang mit den Daten** bei DFG Förderung
    > bisher wenig konkrete Anforderungen bzw. Leitlinien zur Datenmanagementplanung (Struktur, Inhalt …)

F.4 Wesentliche Herausforderungen des Betriebs von FDM-Lösungen? Wie wurden diese bisher adressiert?

Archivierung zur Beweissicherung <> Nachnutzung projektintern + extern

- Archivmitarbeiter brauchen Spezialwissen
  Mischqualifikation: technische, archivarische & sozialwissenschaftlicher Expertise
  - mehrjähriges training on the job & langfristige Perspektive
- Datenmanagement nicht Teil universitärer Ausbildung
  - Wissensvermittlung & Ausbildung in GESIS für Forscher (interne & externe)
- Zentrale Erschließung von SOWI Daten in DE
  - Überlegungen zu gemeinsamen Metadatenkatalog von GESIS / RatSWD
  - Aufbau Datennachweissystem im DFG Projekt da|ra
- Stabile & nachhaltige Daten-/Metadateninfrastruktur
  - GESIS-DAS Beratung von & Kooperation mit nat. & internationalen Forschungsprojekten
  - Förderung Infrastruktur: Aufbau eines Servicezentrum für qualitative Daten (QualiService)
  - Mitgestaltung internationaler Dateninfrastrukturen: CESSDA-ERIC, Data without Boundaries, DAISISH
  - Metadatenentwicklung & Tools im Kontext der Data Documentation Initiative (DDI)