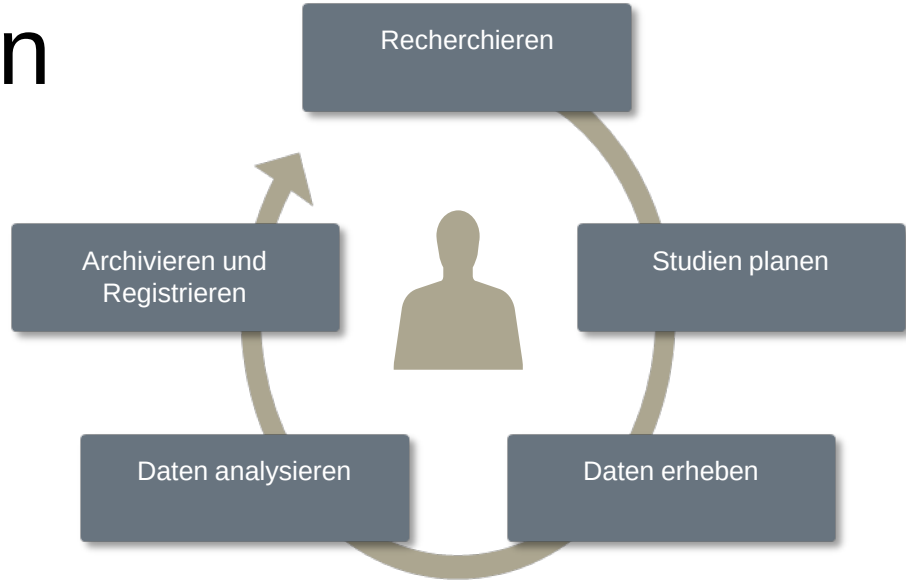
gesis
Leibniz-Institut für Sozialwissenschaften

Das GESIS Datenarchiv für Sozialwissenschaften

Reiner Mauer, GESIS – Leibniz-Institut für Sozialwissenschaften

Workshop „Archivierung sozial- und wirtschaftswissenschaftlicher Datenbestände“
Deutsche Nationalbibliothek, Frankfurt, 15./16. September 2011

GESIS – Leibniz-Institut für Sozialwissenschaften



- Infrastruktureinrichtung für die Sozialwissenschaften
- 250+ Mitarbeiter/innen an vier (ab 11/2011 drei) Standorten Mannheim, Köln, ~~Bonn~~, Berlin)
- „Forschungsdatenzyklus“ dient als Leitbild zur Strukturierung und Verknüpfung der Angebote
- **Archivierung, Dokumentation und Langzeitsicherung** von Forschungsdaten **explizit als Zielsetzung in der Satzung** verankert (§ 2, 2c)

Datenarchiv für Sozialwissenschaften (DAS)

- 1960 als erstes sozialwissenschaftliches Datenarchiv in Europa gegründet (Zentralarchiv für Empirische Sozialforschung, Universität zu Köln)
- 1986: Zentralarchiv wird Mitglied der neu gegründeten GESIS
- 2008: Überführung der drei Teil-Institute der GESIS in ein Institut; seither ist das Archiv eine wissenschaftliche Abteilung der GESIS

➡ 50+ Jahre praktische Datenarchivierung
(Datenbestand war auch durch institutionellen Wandel nie gefährdet bzw. in Frage gestellt)

Leitgedanken bei der Gründung

- Forschungsökonomie: Datenerhebung ist teuer u. zeitaufwändig, Nutzung unausgeschöpfter Forschungspotentiale, Vermeidung von Dopplungen
- Analyse v. sozialem Wandel erfordert Zeitreihen (Daten aus der Vergangenheit)
- Nachvollziehbarkeit von Forschungsergebnissen (Transparenz und Überprüfbarkeit)
- Daten für Methodenforschung, Lehre, zur Vorbereitung neuer Erhebungen …

→ Im Fokus der Arbeit stand immer die Nachnutzung („neue Fragen an alte Daten“)

Aufgaben

- Langzeitarchivierung von Forschungsdaten: Langzeitverfügbarkeit und Interpretierbarkeit sichern
- Sekundäranalysen, Replikationen, Raum- und Zeitvergleiche ermöglichen
- Zugang zu internationalen Forschungsdaten gewährleisten
- Dienstleister für Primärforscher bei der Sicherung, Dokumentation, Aufwertung und Bereitstellung ihrer Daten

Datenbestand

- gegenwärtig ca. 6.100 Datensätze
  (überwiegend Mikrodaten der Umfrageforschung sowie historische Zeitreihen / Aggregatdaten)
- überwiegend Soziologie und Politikwissenschaft
- wichtiger Schwerpunkt sind Daten der interkulturell vergleichenden Sozialforschung (ISSP, EB, EVS, CSES) sowie kontinuierliche nationale Erhebungsprogramme (ALLBUS, GLES, Politbarometer, DeutschlandTrend)
- umfangreiche Bestände zu politischen Einstellungen u. Verhalten, Werten, Jugend, Gesundheit, Mediennutzung ...

Datenbestand (Quellen):

Daten wurden / werden

- … **selbst bzw. unter Beteiligung von GESIS erhoben**:
  z. B. ALLBUS, EVS, ISSP, GLES

- ... **akquiriert** (externe Daten):
  Großteil des Archiv-Bestands, wie z.B. Politbarometer, Eurobarometer, Medialanlyse, Reiseanalyse, viele Einzelstudien)

- ... **entwickelt / produziert / transformiert** (Produktion zeit- u./od. ländervergleichender Datensätze):
  Eurobarometer, EVS, ISSP, Politbarometer, Daten der historischen Statistik (HISTAT)

Datenarchivierung@DAS: Konzeptionell

(1) DAS ist ein **‚Digitales Langzeitarchiv'**

- „Organisation […] die die Verantwortung für den Langzeiterhalt und die Langzeitverfügbarkeit digitaler Objekte sowie für ihre Interpretierbarkeit zum Zwecke der Nutzung durch eine bestimmte Zielgruppe übernommen hat." (Definition nestor Kriterienkatalog, orientiert sich an OAIS)
- Übernahme der Verantwortung für LZA sowie Nutzungs- u. Zielgruppenorientierung explizit in Satzung der GESIS verankert

(2) DAS betreibt **Digital Curation**

- "[...] **maintaining and adding value** [...] **for future and current use**; specifically, the **active management and appraisal of data over the entire life cycle**. [...] builds upon the underlying concepts of **digital preservation** […]. JISC, Digital Preservation briefing paper, 26. Nov. 2006
- relativ neuer Begriff / Konzept: beschreibt aber in aktueller Terminologie am Besten das Verständnis von Datenarchivierung bei GESIS

Datenarchivierung@DAS: Funktional

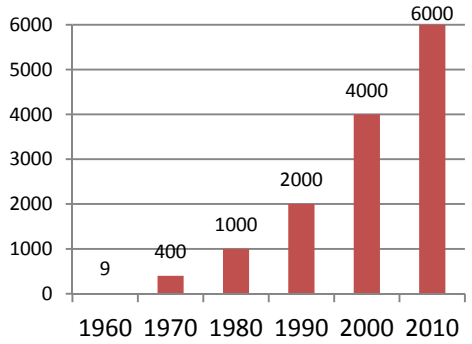
- Akquisition
- Aufnahme ins Archiv (Ingest)
- Datenaufbereitung u. –dokumentation (Standardarchivierung für alle und Added-value für ausgewählte Studien/ Kollektionen)
- Datenregistrierung (da|ra)
- Langzeitarchivierung
- Datenservice (Access): Beratung, Datenbereitstellung (Download, Portale, Online-Analyse, manuelle Bereitstellung).

Akquisition

Kriterien für Auswahl und Bewertung:

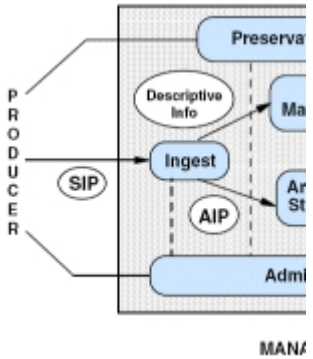
- Disziplinen: Sozialwissenschaften, …
- Datentypen: Umfragedaten, Historische Aggregatdaten …
- Studien erlauben Aussagen über deutsche Bevölkerung od. Teile von ihr
- Beteiligung deutscher Forscher (unabhängig davon, ob Untersuchungsgebiet Deutschland ist oder nicht)
- Daten ermöglichen Raum- / Zeitvergleiche
- Bedeutung für die Forschung
- Komplettierung bestehender Kollektionen

Ungefähre Anzahl von Datensätzen im Archiv



- Wahstudien (1949 ff.)
- Eurobarometer (1970 ff.)
- Political Action I+II (1973/1981)
- Politbarometer (1977 ff.)
- ALLBUS (General Social Survey) (1980 ff.)
- European Values Study (EVS/WVS 1981 ff.)
- International Social Survey Program (1985 ff.)
- …
- USIA “International relations” (1955 ff.)
- Jugendstudien (1962 ff.)
- Tourismusstudien (1971 ff.)
- Mediaanalysen (1972 ff.)
- Wohlfahrtsurvey (series 1978ff.)
- Employment studies (series 1979 ff.)
- Ausländer in Deutschland (series 1984 ff.)
- Familiensurvey (series 1988 ff.)
- …
- DDR Studien (1968 ff.)
- Studien aus Osteuropa (1989 ff.)
- …
- Daten der Historischen Sozialforschung

Ingest / Aufnahme ins Archiv (1)



- Vorbereitung der Archivierung mit Datengeber:
  - Ziel ist es möglichst früh im Lebenszyklus anzusetzen
  - Klärung bzgl. Umfang, Formaten, Aufbereitungs- u. Dokumentationszielen, Zugangsklasse
  - Abschluss Archivierungsvertrag
- Eingangskontrolle SIP und Basis-Aufbereitung (Standardarchivierung)
  - Technische Kontrollen (Formate, Lesbarkeit, Virenfreiheit etc.)
  - Vollständigkeit (Daten, Messinstrumente und Dokumente) u. Nutzbarkeit
  - Übereinstimmung Daten u. Dokumentation
  - Konsistenz der Daten: wild codes, fehlende Werte, question routing (filter), widersprüchliche Merkmalkombinationen
  - Datenschutz
  - ggf. korrigierende Aufbereitungen

Ingest / Aufnahme ins Archiv (2)

- Studienbeschreibung (inhaltliche, methodische, technische Charakteristika)
- Versionierung u. Vergabe eines persistenten Identifikators (DOI)
- Erzeugung AIP (Originale, aufbereite Versionen von Daten und Dokumenten, normalisierte Dateien sowie dazugehörigen Metadaten)
- Erzeugung der für den Service bestimmten Objekte in ein oder mehrere DIPs

Datenaufbereitung / -dokumentation (1)

- Unterscheidung zwischen ***Standardarchivierung*** für alle Studien und darauf aufbauend ***added value*** *Archivierung* für besondere Studienkollektionen
- Standardarchivierung umfasst aufbauend auf Eingangskontrolle u.a.
  - einfache Aufbereitung (Fehlerkorrekturen, Label, fehlende Werte)
  - Vergabe einer DOI
  - Studienbeschreibung (Beschreibung inhaltlicher, methodischer und technischer Charakteristika; DDI-kompatibel)
  - Datenservice / Bereitstellung der Daten

Datenaufbereitung / -dokumentation (2)

Adding value

- hauptsächlich für *ausgewählte* komparative Studien, die kontinuierlich durchgeführt werden bzw. mehrere Länder umfassen
- maßgeschneiderte Unterstützung für große Umfrageprogramme
- häufig ist GESIS bereits an der Datenerhebung beteiligt
- Service u. Forschung dazu i.d.R. durch GESIS FDZs

ALLBUS

EUROBAROMETER

GLES German Longitudinal Election Study

CSES COMPARATIVE STUDY of ELECTORAL SYSTEMS

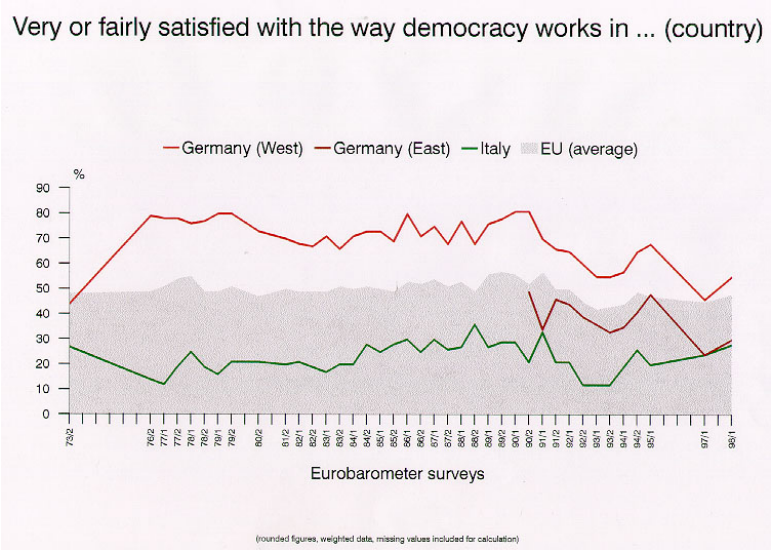
European Values Study

ISSP

Datenaufbereitung / -dokumentation (3)

Adding value: Daten

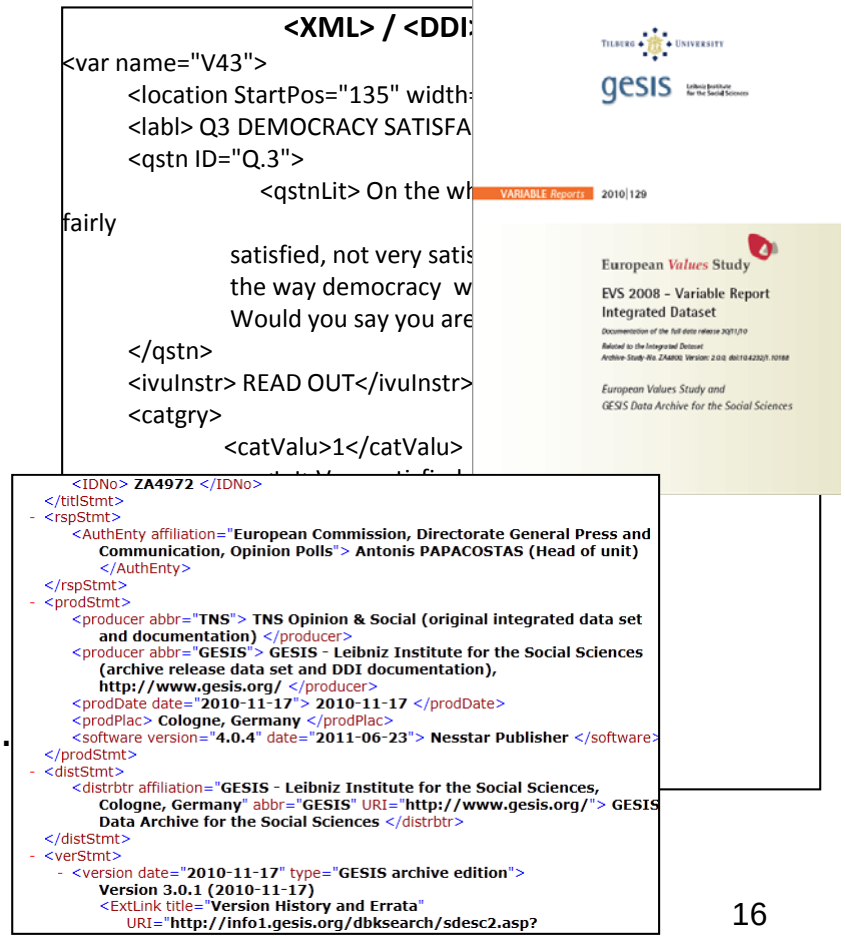
- **Standardisierung**
- **Harmonisierung**
- **Integration / Kumulation**
  (zeit- und/oder ländervergleichend)
- Ergänzung mit **Kontextdaten/ Aggregatdaten**



Datenaufbereitung / -dokumentation (3)

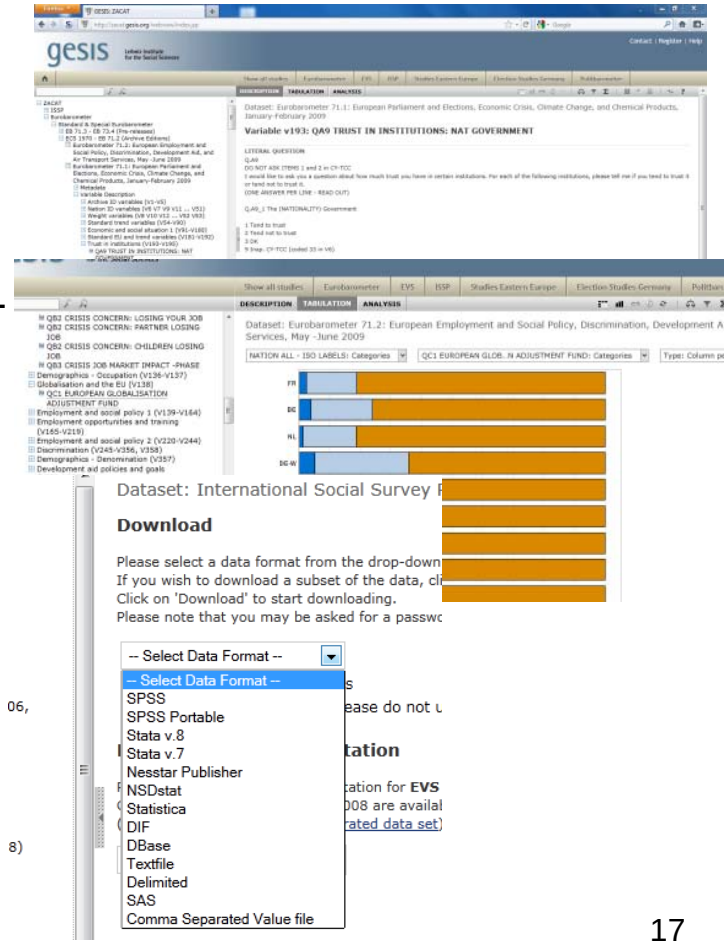
Adding value: Dokumentation

- Umfassende Produktion von strukturierten Metadaten
  (z.B. vollständige u. multilinguale Frage- und Antworttexte, Interviewerанweisung, Anmerkungen zur Datenqualität auf Variablenebene)
- weitere Kontextinformationen
- Codebücher, Datenhandbücher, Variablenreports, Methodenberichte, Dokumentation der Datenaufbereitung….



Access / Datenbereitstellung

- Datenzugang über Online-Portale und (individuelle) Bereitstellung auf Datenträgern oder per ftp
- (Meta)Datenportale: ZACAT, HISTAT, CESSDA-Portal, sowiport, DBK (Retrieval, Download, Online-Analyse)
- DAS vermittelt Zugang zu Datenbeständen ausländischer Archive (ICPSR, CESSDA, IFDO)
- 30.000+ Datenweitergaben in 2010
- Weit überwiegend akademische Nutzung: Lehre und Forschung
- Je nach Angebot 30%-70% internationale Nutzer

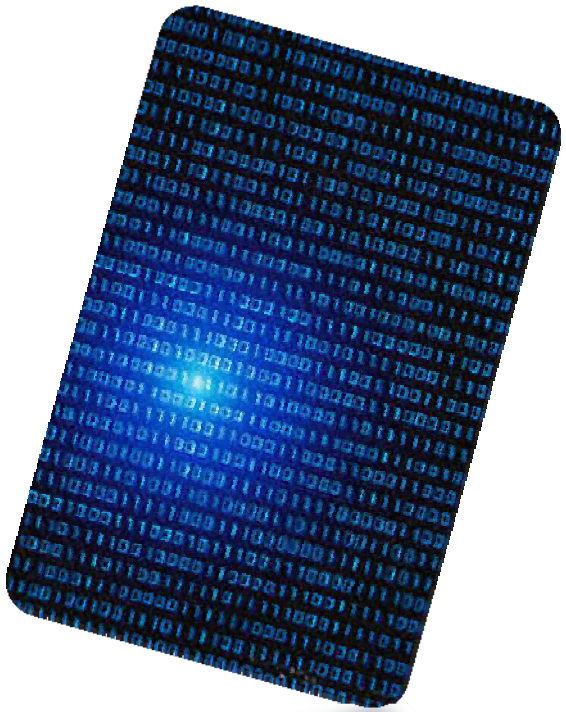


Langzeitarchivierung: Archivspeicher

- Archivierung im engeren Sinne erfolgt durch die Überführung der zum AIP gehörenden Objekte in den zentralen Archivspeicher (STAR = Studienarchiv)
- Organisation des Archivspeichers:
  - dateibasierte Verzeichnisstruktur
  - alle zu einer Studie gehörenden digitalen Objekte (AIP und DIP) werden dort so abgelegt, dass sie den Archiv-Lebenszyklus einer Studie reflektieren
  - Dateien werden nach bestimmten Regeln in definierte Verzeichnisse abgelegt u. nach einheitlichen Schema benannt
  - Restriktive Zugangsrechte

Langzeitarchivierung: Substanzerhaltung

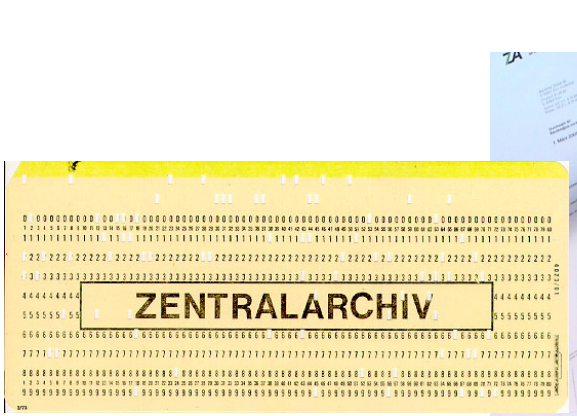
- physischer Schutz vor Ort
- räumlich getrennte und redundante Datenhaltung
- Diversivität eingesetzter Speichertechnik
- regelmäßige Medienmigration / Refreshment

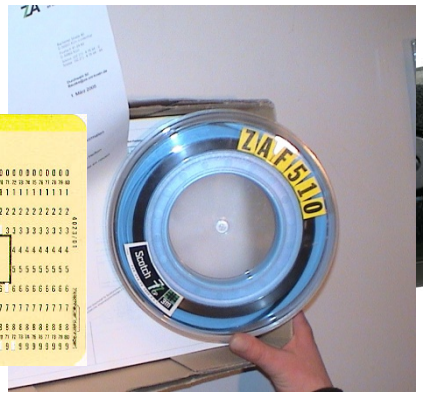


Langzeitarchivierung: Erhalt von Nutzbarkeit und Interpretierbarkeit

- Bedrohung der digitalen Bestände durch technischen Fortschritt ist eine Konstante in der 50jährigen Archivarbeit
- Maßnahmen, die auf Erhalt des bitstreams abzielen nicht ausreichend

→ Erhalt wird hauptsächlich durch Migrationsstrategien erreicht (zur Überbrückung auch Emulation bzw. Virtualisierung)





Langzeitarchivierung: Erhalt von Nutzbarkeit und Interpretierbarkeit (2)

Maßnahmen:

- Verfolgen der Entwicklung von Speichertechnik, Medien, Software und **insbesondere damit verbundener Dateiformate**
- Überführung d. Objekte in definierte und standardisierte Formate (Ingest)
  - erleichtert Monitoring
  - deutlich verringerter Ressourcenbedarf bei Migration
  - Auswahl guter Formate senkt Migrationsbedarf
- reine Datenträgermigrationen finden kontinuierlich statt.
- **Formatmigrationen**, nur wenn Gefahr der Beeinträchtigung der Nutzbarkeit oder wenn mit der Migration so große Vorteile für die Nutzung oder die Archivarbeit einhergehen, dass der Aufwand zu rechtfertigen ist

Ausblick

Herausforderungen und Chancen

- Dateninfrastrukturen werden verteilter
- Bedeutung komplexer Forschungsdesigns und neuer Datenformen wächst
- Anforderungen an Forscher bzgl. Datenmanagement wachsen

Weiterentwicklung der Angebote

- DAS bietet zwar umfassenden Archivierungsservice, aber sehr starker Fokus auf Umfragedaten
- Stärkere Modularisierung des Archivierungsservices, um bedarfsgerechte Dienste bieten zu können:
  - Langzeitarchivierung (projektspezifisch, bspw. als Hintergrunddienst für Datenzentren)
  - Datenservice: Datenzugang u. Nutzerservice (projektspezifisch)
  - Aufbereitung u. Dokumentation (projektspezifisch, bspw . große Erhebungsprogramme)
  - Datenregistrierung
  - Zentrale Datennachweisdienste (also auch für Daten außerhalb der GESIS)
  - „Selbstarchvierung“ für Daten (niedrigschwelliger Weg ins Archiv)