gesis
Leibniz-Institut für Sozialwissenschaften

<ddi>

Metadaten und die Data Documentation Inititive (DDI)

Wolfgang Zenk-Möltgen
GESIS – Leibniz Institut für Sozialwissenschaften
Wolfgang.Zenk-Moeltgen@gesis.org

Workshop „Archivierung sozial- und wirtschaftswissenschaftlicher Datenbestände“
Frankfurt/Main, 15.-16.09.2011

Themen

- Über DDI
- Grundlegende DDI Metadaten
- DDI im GESIS Datenarchiv
- Standardisierung

Über DDI

DDI – Data Documentation Initiative

Eine Initiative um einen internationalen Standard zur Beschreibung sozialwissenschaftlicher Daten zu definieren.
Ein XML-Format, das sowohl mensch- als auch maschinenlesbar ist.
Unterstützung des Forschungsdaten-Lebenszyklus.
DDI Metadaten beziehen sich auf Studienkonzeption, Datenerhebung, Datenbearbeitung und -auswertung, sowie Sekundärnutzung und Archivierung.

http://www.ddialliance.org/

Entstehung von DDI

- Konzeption und Definition der Ziele kamen aus der Welt sozialwissenschaftlicher Datenarchive
- 1995 als Projekt finanziert, gestartet und organisiert vom ICPSR
- 2003 – Gründung der DDI Alliance
  - Basiert auf Mitgliedschaft von Institutionen
  - Formalisierte Prozesse zur Weiterentwicklung

Mitglieder von DDI

- Gründungsmitglieder
  - Sozialwissenschaftliche Datenarchive
  - Produzenten von Statistikdaten
- Weitere Mitglieder
  - Forschungsdatenzentren
  - Datenerhebungsinstitutionen
  - Kommerzielle Organisationen

University of Alberta, Canada
Australian Bureau of Statistics (ABS)
Australian Social Science Data Archive (ASSDA)
University of California, Berkeley -- Computer-Assisted Survey Methods Program and UCDATA
University of California, California Digital Library
Centro De Investigaciones Sociologicas (CIS), Spain
CEPS/INSTEAD -- Luxembourg
Cornell University (CISER)
Danish Data Archive
Data Archiving and Networked Services (DANS), The Netherlands
Finnish Social Science Data Archive
German Socio-Economic Panel Study (SOEP)
GESIS - Leibniz Institute for the Social Sciences
University of Guelph
Institute for Quantitative Social Science (IQSS) at Harvard University
Institute for the Study of Labor (IZA)
Inter-university Consortium for Political and Social Research (ICPSR)
Massachusetts Institute of Technology (MIT)
University of Minnesota, Minnesota Population Center
National Opinion Research Center (NORC)
Norwegian Social Science Data Service (NSD)
Open Data Foundation
Princeton University
Research Data Centre of the German Federal Employment Agency, Institute for Employment Research (IAB)
Roper Center
Stanford University
Survey Research Operations, University of Michigan
Swedish National Data Service (SND)
Swiss Foundation for Research in Social Sciences (FORS)
United Kingdom Data Archive
University of Toronto
University of Wisconsin
U.S. Bureau of Labor Statistics (Associate Member)
World Bank, Development Data Group (DECDG)
Yale University

DDI wird in aller Welt angewendet

DDI Spezifikationen

- 2000 – DDI 1.0
  - Dokumentation einfacher Unfragen, nur Mikrodaten
- 2003 – DDI 2.0 and 2.1 “DDI-Codebook”
  - Erweiterung um Aggregatdaten
  - Unterstützung für geographische Elemente
- 2008 – DDI 3.0 “DDI-Lifecycle”
  - Lebenszyklus-Modell: Vom “Codebuch”/Variablen-zentrierten Modell zur Erfassung des Daten-Lebenszyklus
  - Blick auf Metadaten-Erzeugung und Wiederverwendung
  - “Machine-actionable” zur Unterstützung von Programmierung
  - CAI Instrumente werden durch eine Erweiterung der Fragebogendokumentation unterstützt
  - Unterstützung für Datenreihen (Längsschnitt-Umfragen, Panel Studien, etc.)
  - Vergleichbarkeit “by design” und “ex-post” möglich
  - Verbesserte Unterstützung zur Beschreibung komplexer Datensätze
- 2009 – DDI 3.1
  - Fehlerkorrekturen
  - Einführung einer URN Struktur, um dauerhafte Identifikatoren aller “identifiable” Elemente zu erhalten
- 2011 (angekündigt) – DDI 2.5
  - Erleichterung der Migration nach DDI 3.x durch Einführung der Notwendigen Elemente
- 2011 (angekündigt) – DDI 3.2
  - Einführung eines “DataItem” zur Wiederverwendung
  - Klärung von “RecordRelationship” weiteren Konsistenz-Issues, etwa bei MissingValues
  - Überarbeitung der URN-Struktur zur Einführung eines verteilten Resolving-Mechanismus
  - Verbesserung der Verwendung kontrollierter Vokabulare

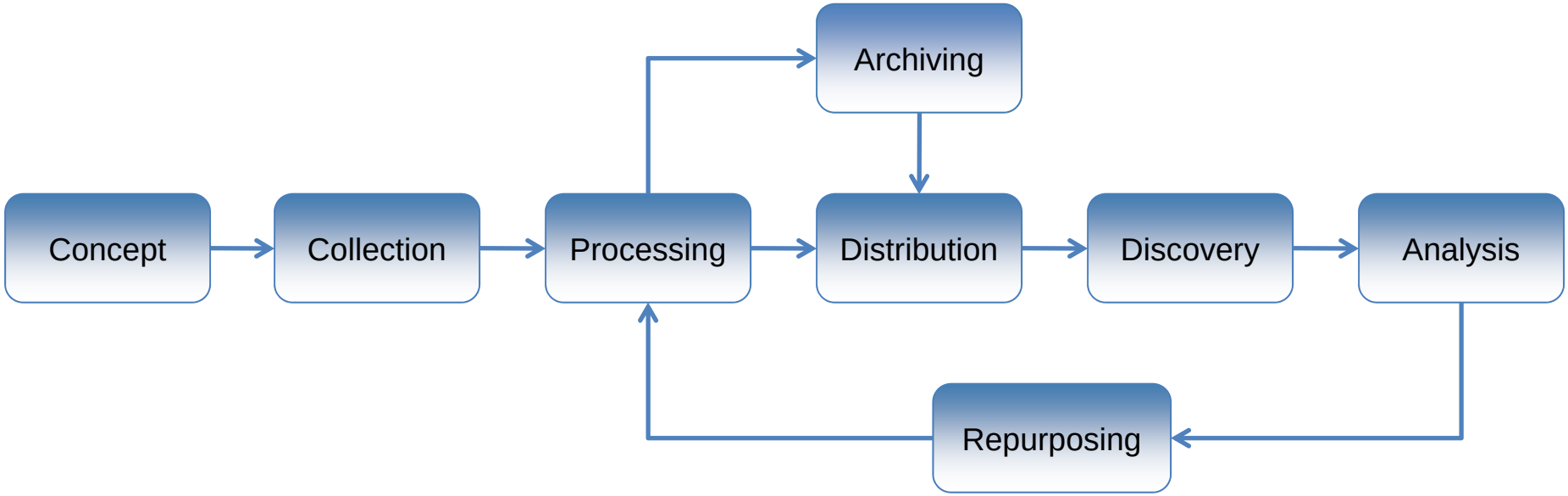
Grundlegende DDI Metadaten

- Dokumentbeschreibung
  - Titel, Autoren, Publikation
- Studienbeschreibung
  - Inhalte, Autoren, Institutionen
  - Zeitliche und geographische Angaben
  - Methoden, Grundgesamtheit, Stichprobe
  - Literaturhinweise
- Variablenbeschreibung
  - Namen, Typ und Labels
  - Fragen und Antworten
  - Codes und Häufigkeiten
  - Intervieweranweisungen, Filter
  - Hinweise zur Codierung oder Berechnung
- Dateibeschreibung
  - Anzahl der Variablen, Fälle
  - Namen, Formate und Versionen

DDI 3 Konzepte

- Lebenszyklus-Konzept
- Wiederbenutzbare Dokumentation
  - Module
  - Maintainables, versionables, identifiables
  - Scheme-basiert (pflegbare Listen)
- Beziehungen zu anderen Standards
- Kontrollierte Vokabulare

Forschungsdaten-Lebenszyklus



DDI 3.1 Module

Sind Gruppen von zusammengehörigen Dokumentationselementen

Manche beziehen sich auf das Lebenszyklusmodell, andere sind technisch gruppiert

- Archive module
- Comparative module
- Conceptual components module
- Data collection module
- Dataset module
- Dublin Core Elements module
- DDI profile module
- Grouping module
- Instance module
- Logical product module
- Physical data product module
  - (plus inline n-cube, normal n-cube, tabular n-cube module and proprietary module)
- Physical instance module
- Reusable module
- Study unit module

Benutzung der DDI 3 Module

Study Unit
- Identification
- Coverage
  - Topical
  - Temporal
  - Spatial
- Conceptual Components
  - Universe
  - Concept
  - Representation (optional replication)
- Purpose, Abstract, Proposal, Funding

Data Collection
- Methodology
- Question Scheme
  - Question
  - Response domain
- Instrument
  - using Control Construct Scheme
- Coding Instructions
  - question to raw data
  - raw data to public file
- Interviewer Instructions

Logical Product
- Category Schemes
- Coding Schemes
- Variables
- NCubes
- Variable and NCube Groups
- Data Relationships

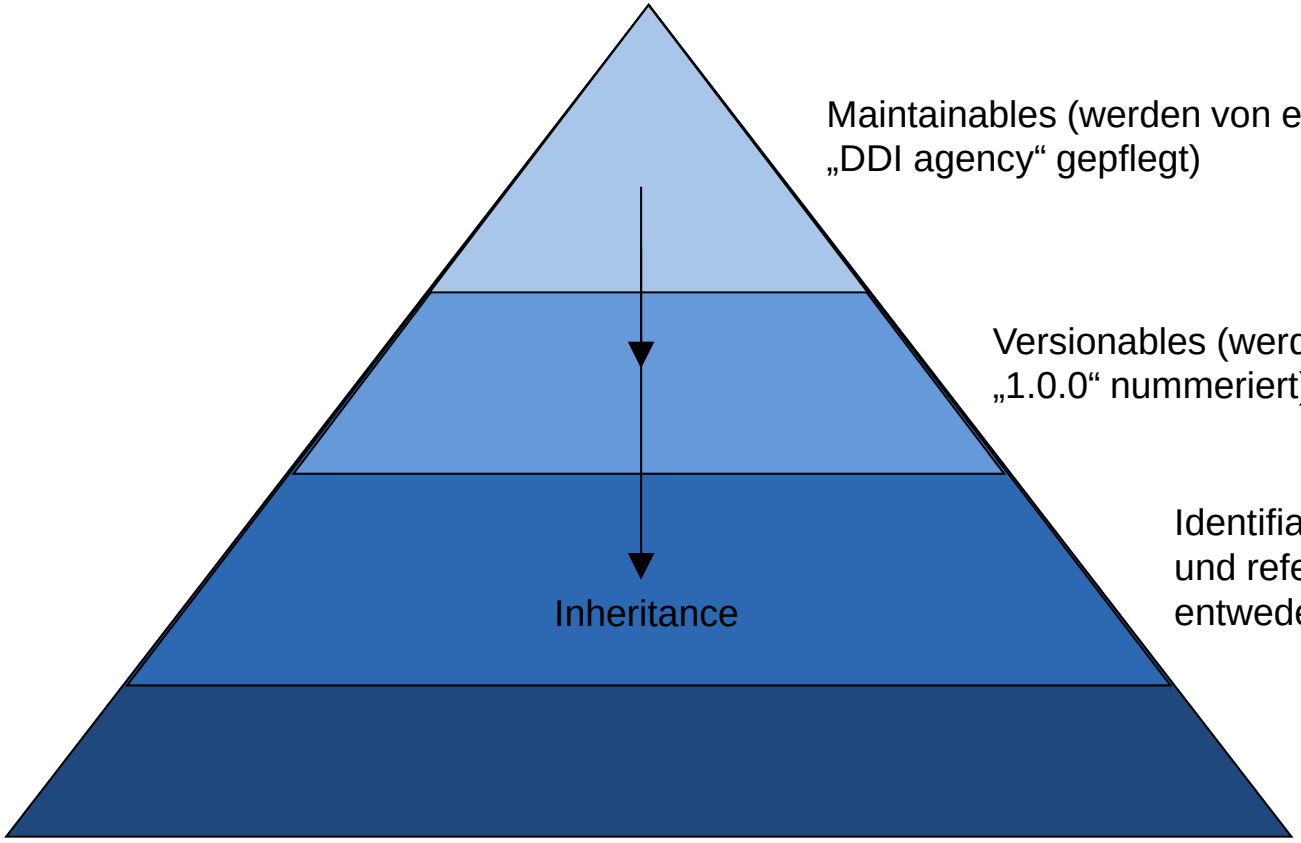
Archive
- Organization or individual which has control over the metadata
- Lifecycle events
- Archive specific information

Physical Data Structure
- Links to Data Relationships
- Links to Variable or NCube Coordinate
- Description of physical storage structure
  - in-line, fixed, delimited or proprietary

Physical Instance
- One-to-one relationship with a data file
- Coverage constraints
- Variable and category statistics

etc…

Maintainables, Versionables, Identifiables



DDI Schemes

Schemes = Listen von Elementen eines Typs
Beispiele

- archive
  - OrganizationScheme
- datacollection
  - QuestionScheme
  - ControlConstructScheme
  - InterviewerInstructionScheme
- conceptualcomponent
  - ConceptScheme
  - UniverseScheme
  - GeographicStructureScheme
  - GeographicLocationScheme
- logicalproduct
  - CategoryScheme
  - CodeScheme
  - VariableScheme
  - NCubeScheme
- physicaldataproduct
  - PhysicalStructureScheme
  - RecordLayoutScheme

DDI und andere Standards

- Dublin Core
  - Grundlegende bibliographische Zitationsinformationen
  - Grundlegende Information zum Bestand und Format
- METS
  - Beschreibende Information zur Verwaltung digitaler Objekte
  - Spezielle Strukturen für fachspezifische Metadaten
- OAIS
  - Referenzmodell für den Archiv-Lebenszyklus
- PREMIS
  - unterstützt und dokumentiert den Prozess der digitalen Sicherung
- ISO 19115 – Geography (FGDC)
  - Metadatenstruktur für geographische Objektdaten wie shape, boundary oder map image Dateien und ihre Attribute
- ISO/IEC 11179
  - Internationaler Standard, um Metadaten in einer “Metadata Registry” zu repräsentieren
  - Besteht aus Hierarchie von “Konzepten” mit zugeordneten Eigenschaften
- SDMX
  - Austausch von statistischen Informationen (Zeitreihen/Indikatoren)
  - Unterstützt Metadatenerfassung ebenso wie die Einrichtung von Registries

Kontrollierte Vokabulare

- Sind nicht Teil des DDI Standards
- Empfehlungen zu:

- Beispiel:
  TimeMethod kann enthalten
  - Longitudinal (Cohort or Trend)
  - Panel (Continuous or Interval)
  - TimeSeries (Continuous or Discrete)
  - CrossSectional
  - CrossSectionalAdHocFollowUp
  - Other

LifeCycleEventType

CommonalityTypeCoded

TimeMethod

ResponseUnit

AggregationMethodsType

DataType

SoftwarePackage

CharacterSet

CategoryStatistic

SummaryStatistic

Date@Calendar

AnalysisUnit

Contributor@Role

Publisher@Role

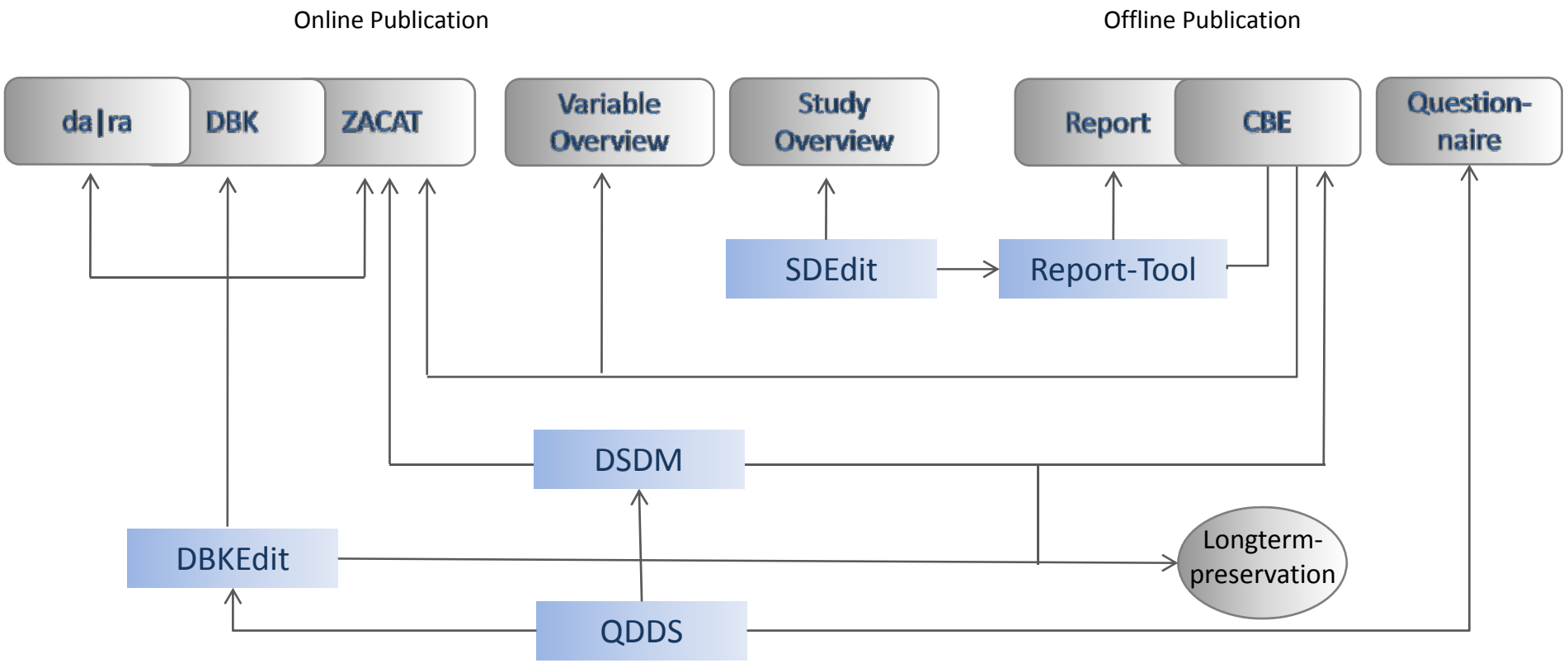
Identifikation in DDI 3

- Zwei Möglichkeiten um ein Element zu identifizieren:
  - Festlegen des <ID> Tags
    - Agency und Version werden geerbt
  - Benutzung der speziell strukturierten URN
    - Agency und Version müssen angegeben werden
- Der Ansatz der strukturierten URN wird empfohlen
- Diese IDs/URNs können referenziert werden
- Beide Ansätze brauchen einen Resolver Service, der zu den Namen die Orte ermittelt, um effektive Re-Analysen möglich zu machen
- DDI Alliance arbeitet zur Zeit daran, auf einem DNS (Domain Name System) basierten Infrastruktur-Ansatz
- Registrierung einer Agency ist bereits möglich

DDI im GESIS Datenarchiv

- Workflow für Daten- und Metadatenbearbeitung
- DDI 2 Metadaten
- Unterstützung von
  - Dokumentation
  - Langfristige Sicherung
  - Recherche und Datenservice für Sekundärnutzer
  - DOI Registrierung (da|ra und DataCite)
- DDI 3 für spezielle Anwendungen
  - Enhanced Publications (Verbindung von Publikation zu den benutzten Daten)
  - Projekt STARDAT (Integration der Archiv-Tools und DDI3)

Workflow



DBK	Data Catalogue	DSDM	Dataset Documentation Manager
ZACAT	Online Study Catalogue	QDDS	Questionnaire Development Documentation System
DBKEdit	Data Catalogue Edit-Tool	CBE	CodebookExplorer
SDEdit	Editing-Tool for Study Method Reports	da|ra	Registration Agency

Beispiel Datenbestandskatalog

- Studienbeschreibungen
  - Generelles Schema zur Beschreibung archivierter Daten seit den 1960er Jahren
  - Weiterentwicklung im internationalen Archivkontext und der DDI Alliance
  - Metadaten-Produktion zur Publikation in verschiedenen Retrieval- und Distributionsplattformen
- Einführung einer Versionshistorie der Daten
  - Dokumentation von Errata und der Korrektur-History
  - Erlaubt einfaches Zitieren der Daten
- Weitere Standardisierung der Angaben
  - Untersuchungsgebiet nach ISO3166-1 und ISO3166-2
  - Erhebungszeiträume im Format TT.MM.JJJJ (inkl. Zeiträume)
  - Literaturangaben

Errata & Versionen

Errata in aktueller Version	2011-3-15	v1-v5; v106; v106_cs; v108_cs; v136 - v147; v308; v322; v353m_pp; v355; v368b_N3; v368b_N2; v368b_N1; v368b_cc; v372; v374b; weight_c	Please download patch and documentation for correcting errata as of 2011-03-15 in EVS 2008 Integrated Dataset (v. 2.0.0): ZA4800_v2-0-0_patch_1.zip; ZA4800_v2-0-0_p1_readme.doc
	2011-3-15	v106_cs v108_cs v264 v265 v305b v307b v310b v312b v343b v368b_CC v336_cs v344_cs v355_csv353W_cs v353M_cs v353Y_cs v368b_N3 v371b_N3 v368b_N2 v371b_N2 v368b_N1 v371b_N1 v376	Correction of value labels with country specific characters: Please download the Unicode patch_2 for correcting the labels in the Integrated Dataset (v. 2.0.0): ZA4800_v2-0-0_patch_2.zip
	2011-3-15	v1 to v5	Correction of the order of variables v1 to v5 in the Swedish data set: v1=v2, v2=v3, v3=v4, v4=v5, v5=v1.
	2011-3-15	v106	In the Norwegian data set hindu is coded as ´5: muslim´, but should be ´6: hindu´.
	2011-3-15	v106_cs, v108_cs	Correction of value label of country specific code 498096 and addition of missing value label of code 499001.
	2011-3-15	v136 to v147	Change of value labels of v136 to v147 into 1 "very important" 2 "rather important" 3 "not very important".
	2011-3-15	v284 to v294	Notification of deviant question wording of Q83 and Q84 . The phrase “feel concerned about” has been translated differently in several field questionnaires, for instance, in some cases it has been translated into “worried about”, in other cases as “involved in”.
	2011-3-15	v308	Illogical answer pattern: In 27 cases (AZ, HR, NCY, FR, DE, LV, LU, MD, SK, SI, ES, UA) is the

Versionshistorie von Daten

- Der GESIS Datenkatalog enthält Links zum Datenzugang
- Einführung einer einheitlichen Versionierung der Daten seit April 2010
- Beginn mit Version 1.0.0 und Erhöhung von Major, Minor, Revision je nach Änderung im Datensatz
- Zu jeder publizierten Version wird eine DOI erzeugt
- Führt zu Transparenz im Laufe der Datenbearbeitung
- Zitation von Datensätzen enthält daher die genaue Version zur Erleichterung von Replikationsanalysen
- Dokumentation der kompletten History in DDI steht noch aus

Standardisierung

- Standardisierte Dokumentation
  - erlaubt den einfachen Austausch
  - führt zur einfacheren Übernahme in neue Systeme/Anwendungen
  - enthält klarere Bedeutungen
  - ist für die langfristige Sicherung unerlässlich

  - muss in der Community erreicht werden
  - erfordert einen höheren Aufwand
  - ist ein dauernder Prozess

Metadata powered by DDI

Vielen Dank für Ihre Aufmerksamkeit!