Mikrodaten-Informationssystem MISSY: Metadaten zu den Mikrozensus Scientific Use Files (Abschlussbericht MISSY II)

Bohr, Jeanette; Hopt, Oliver; Lengerer, Andrea; Schroedter, Julia H.; Wira-Alam, Andias

Veröffentlichungsversion / Published Version
Abschlussbericht / final report

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
GESIS - Leibniz-Institut für Sozialwissenschaften

gesis Leibniz-Institut für Sozialwissenschaften

TECHNICAL *Reports* 2010|07

Mikrodaten-Informationssystem MISSY: Metadaten zu den Mikrozensus Scientific Use Files (Abschlussbericht MISSY II)

Jeanette Bohr, Oliver Hopt, Andrea Lengerer, Julia Schroedter, Andias Wira-Alam

GESIS-Technical Reports 2010|07

Mikrodaten-Informationssystem MISSY: Metadaten zu den Mikrozensus Scientific Use Files

(Abschlussbericht MISSY II)

Jeanette Bohr, Oliver Hopt, Andrea Lengerer,
Julia Schroedter, Andias Wira-Alam

GESIS – Leibniz-Institut für Sozialwissenschaften 2010

GESIS-Technical Reports

GESIS – Leibniz-Institut für Sozialwissenschaften
Postfach 12 21 55
68072 Mannheim
Telefon: (0621) 1246 - 261
Telefax: (0621) 1246 - 100
E-Mail: jeanette.bohr@gesis.org

ISSN: 1868-9043 (Print)
ISSN: 1868-9051 (Online)
Herausgeber,
Druck und Vertrieb: GESIS - Leibniz-Institut für Sozialwissenschaften
Lennéstraße 30, 53113 Bonn

Anschlussvorhaben Forschungsverbund Datenzentren. Verbesserung des Zugangs der Wissenschaft zu Mikrodaten Teilprojekt: Pilotprojekt zum Aufbau eines Servicezentrums für Mikrodaten der GESIS. Mikrodaten-Informationssystem MISSY

BMBF Projekt
Förderkennzeichen: 01UW0707

Gesamtbewilligungszeitraum 01.01.2008 - 31.12.2009

Abschlussbericht

Jeanette Bohr, Oliver Hopt, Andrea Lengerer, Julia Schroedter, Andias Wira-Alam

April 2010

Vorhabenleiter: Prof. Dr. Christof Wolf, Prof. Dr. York Sure

Ausführende Stelle:

GESIS
Postfach 12 21 55
68072 Mannheim

GESIS
Lennéstraße 30
53113 Bonn

Inhaltsverzeichnis

I Kurzdarstellung

Dieser Bericht informiert über die durchgeführten Arbeiten im Rahmen des Projekts MISSY II (Projektlaufzeit 01.01.2008 - 31.12.2009). Es handelt sich dabei um ein Anschlussvorhaben zum Aufbau des Mikrodaten-Informationssystems MISSY, welches den inhaltlichen Ausbau von MISSY um die Metadaten der Mikrozensus Scientific Use Files der Jahre 1973 bis 2006 sowie die informationstechnologische Neuimplementierung des Systems beinhaltet.

1 Aufgabenstellung

Im Rahmen des Forschungsverbundes „Datenservicezentrum – Verbesserung des Zugangs der Wissenschaft zu Mikrodaten" wurde im Zeitraum vom 01.07.2003 bis 31.12.2006 im Teilprojekt zum Aufbau eines Servicezentrums für Mikrodaten der GESIS ein Prototyp des Mikrodaten-Informationssystems MISSY aufgebaut. Die Weiterfinanzierung des Projekts (MISSY II) sollte den Ausbau des Prototypen zu einem umfassenden und zukünftig nutzbaren Metadatensystem gewährleisten.

Ziel von MISSY ist es, den Zeit- und Arbeitsaufwand für Wissenschaftler, die mit den Scientific Use Files (SUF) des Mikrozensus arbeiten, zu verringern, und deren Analysen durch Bereitstellung relevanter – bis dato z. T. schwer zugänglicher – Informationen zu bereichern. Dabei ist die Dokumentation aktueller Datensätze nicht ausreichend, vielmehr verweist die Nutzung älterer Mikrozensen in der Forschung darauf, dass die älteren Jahrgänge ebenfalls dokumentiert werden müssen. Die Weiterfinanzierung des Projekts MISSY hat entsprechend die volle Implementierung des System zur Aufgabe und umfasst die Einbindung von Metadaten zu allen derzeit[1] verfügbaren Scientific Use Files des Mikrozensus-Grundfiles von 1973 bis 2006. Zudem wurde MISSY um Serviceangebote zur Datenaufbereitung und -analyse erweitert und die Struktur im Hinblick auf eine Integration weiterer Datensätze (z. B. Mikrozensus-Panelfile) ausgebaut. Neben dem inhaltlichen Ausbau des Systems ist als weiterer Aufgabenschwerpunkt die informationstechnologische Neuimplementierung zu nennen. Dabei lag der Fokus auf der Entwicklung eines Systems, das einen langfristigen Einsatz gewährleistet. Wichtig war daher, dass es auf der aktuellen Version eines etablierten Datenstandards (DDI 3.0) beruht, in seiner Benutzungsoberfläche so flexibel wie möglich gestaltet ist, nicht auf die Dokumentation des Mikrozensus beschränkt bleibt und möglichst leicht an zukünftige Entwicklungen im Bereich Datenstandards angepasst werden kann.

2 Voraussetzungen

Beim Ausbau und der Weiterentwicklung von MISSY konnte einerseits auf die inhaltlichen und konzeptionellen Vorarbeiten aus dem Pilotprojekt MISSY I und den darin entwickelten Prototypen des Informationssystems, andererseits auf die Ergebnisse einer nach Abschluss des Pilotprojektes durchgeführten Nutzerstudie zurückgegriffen werden. Im Rahmen der Evaluierung des German Microdata Labs als Servicezentrum für Mikrodaten der GESIS im Jahr 2006 wurde der Nutzen von MISSY für die Wis-

[1] Stand vom Dezember 2009.

senschaft bestätigt und die Weiterförderung von MISSY infolgedessen auch vom Rat für Sozial- und Wirtschaftsdaten empfohlen.

Die Grundlage für die inhaltliche Aufbereitung von Metadaten bilden neben der Berücksichtigung internationaler Metadatenstandards die jahrelangen Erfahrungen des German Microdata Labs im Bereich amtlicher Mikrodaten, insbesondere mit dem Mikrozensus. Im Rahmen seiner Beratungsfunktion für die Wissenschaft hat das German Microdata Lab einen guten Überblick über die Fragen und Probleme, auf die die Wissenschaftler im Umgang mit dem Mikrozensus stoßen, und auch durch die eigene Forschungstätigkeit seiner wissenschaftlichen Mitarbeiter verfügt das German Microdata Lab über sozialwissenschaftliches Know-how und Erfahrung im Umgang mit den Daten des Mikrozensus.

Für eine nachhaltige Konzeption im Bereich der Re-Implementierung waren verschiedene Erfahrungen der GESIS Abteilung „Informationelle Prozesse in den Sozialwissenschaften" (IPS) hilfreich. So ist im Rahmen des DFG-Projekts QDDS bereits ein Editor für einen anderen Ausschnitt von DDI-konformen Daten implementiert worden. Die Architektur dieses Editors und einige zentrale Komponenten konnten in das Projekt mit einfließen. Außerdem kann die IPS auf umfangreiche Erfahrungen im Zusammenhang mit der Planung und Umsetzung von Informationssystemen zurückgreifen.

3 Planung und Ablauf

Die Planung sah zunächst die Entwicklung eines neuen Erfassungswerkzeugs für DDI-konforme Metadaten vor, um die bislang eingesetzte Lösung auf der Basis von Microsoft InfoPath abzulösen. Während der Entwicklungsphase des neuen Editors wurde die Erfassung der variablenbezogenen Metadaten zunächst noch über das aus der ersten Projektphase vorhandene Werkzeug (Microsoft Infopath) weitergeführt und nach Fertigstellung in den neuen Editor konvertiert. Die Eingabe aller weiteren Metadaten wurde anschließend über das neue Erfassungswerkzeug durchgeführt. Für den Zugang zu den Informationen auf Variablenebene wurden die verschiedenen Nutzerzugänge mittels DBClear realisiert. Parallel dazu wurden über das Content Managment System Typo3 die MISSY-Webseiten erstellt und PDF-Dokumente (Fragebögen etc.) sowie Syntaxdateien eingebunden.

4 Wissenschaftlicher und technischer Stand, an den angeknüpft wird

Die Entwicklung von MISSY II knüpft inhaltlich und konzeptionell eng an das in der ersten Projektphase entwickelte System MISSY I an und berücksichtigt neben den Ergebnisse aus einer MISSY-Nutzerstudie auch die Anregungen aus der Begutachtung des Projektvorhabens im Auftrag des Rates für Sozial- und Wirtschaftsdaten.

Die grundlegende Konzeption des Systems, die in der ersten Projektphase entwickelt wurde, orientierte sich fachlich insbesondere am Informationssystem des British Household Panel Survey, da deutschsprachige Informationssysteme, die über einen ähnlichen Umfang und ein vergleichbares Funktionsreichtum wie MISSY verfügen, in den Wirtschafts- und Sozialwissenschaften – bis auf das für den Paneldatensatz SOEP angelegte SOEP-info – bislang kaum zu finden sind. Struktur und Umfang des Metadatenangebots in MISSY orientieren sich an dem Datendokumentationsstandard der Data Documentation

Initiative (DDI) sowie dem sich in der täglichen Arbeit des GML herauskristallisierenden Dokumentationsbedarf.

Auf der technischen Seite war bereits eine Architektur für einen Metadateneditor vorhanden, welche direkt auf die Anforderungen von MISSY übertragbar war. Zudem ist im Rahmen der Unterstützung der sozialwissenschaftlichen Fachinformation ein Softwaresystem namens DBClear entstanden, das große Teile eines Informationssystems bereits implementiert und sehr variabel auf ein spezifisches Informationsangebot angepasst werden kann. Dies gilt für das verwendete Datenschema, für die verwendeten Vokabularien sowie für die Umsetzung der Informationsabfrage.

5 Zusammenarbeit mit anderen Stellen

Inhaltlich erfolgte die Weiterentwicklung von MISSY – wie schon im Pilotprojekt – im kontinuierlichen Austausch mit der Fachabteilung Mikrozensus des Statistischen Bundesamtes (Gruppe VIII C) in Bonn, auf dessen Dokumente bei der Erstellung von Metadaten zurückgegriffen werden konnte.

II Eingehende Darstellung

1 Inhaltliche Arbeiten

1.1 Inhaltliche Konzeption eines Erfassungswerkzeuges

Da sich im Pilotprojekt herausgestellt hatte, dass existierende Datenerfassungssysteme Defizite in Bezug auf Flexibilität und Leistungsfähigkeit aufweisen, wurde bei Projektbeginn zunächst ein Erfassungstool entwickelt, über welches die Eingabe von Informationen in das neue Datenbanksystem (MISSY II) in standardisierter Weise erfolgt. Bei der inhaltlichen Konzeption des Tools wurde das Ziel verfolgt, ein System zu entwickeln, das den komplizierten und umfangreichen Metadaten des Mikrozensus optimal gerecht wird und gleichzeitig eine effiziente und sichere Informationserfassung ermöglicht.

Das Tool gewährleistet zum einen, dass die Detailinformationen der verschiedenen Mikrozensus-Erhebungsjahre an einer zentralen Stelle eingegeben, gesammelt und gespeichert werden. Darüber hinaus erfüllt das Erfassungstool die Aufgabe, die eingegebenen Informationen gewissermaßen „automatisch" in DDI 3.0 zu konvertieren.

Das Erfassungstool wurde zudem im Hinblick auf die Pflege der Datenbank, darunter auch die zukünftige Aufnahme weiterer Mikrozensus-Erhebungszeitpunkte sowie sonstiger Datensätze, erstellt. Eine genauere Darstellung der Arbeitsabläufe zum Einrichten neuer Erhebungszeiträume findet sich in Kapitel 2.1. Im Folgenden werden die Funktionen des so genannten MISSY-Editors im Einzelnen erläutert.

Eingabe der Studieninformation

Der Zugriff auf und die Eingabe in den MISSY-Editor ist über eine autorisierte Anmeldung möglich. Die Eingangsseite (mit der Karteikarte „Studie") informiert zunächst über den geöffneten Datensatz und den ausgewählten Erhebungszeitpunkt. Hier wird auch der Wechsel zwischen verschiedenen Studien und/oder Erhebungsjahren vorgenommen.

Über die Auswahl der Karteikarte „Variable" gelangt man zur Eingabemaske der Detailinformationen des jeweils ausgewählten Datensatzes und Erhebungsjahres.

Eingabe der Detailinformationen

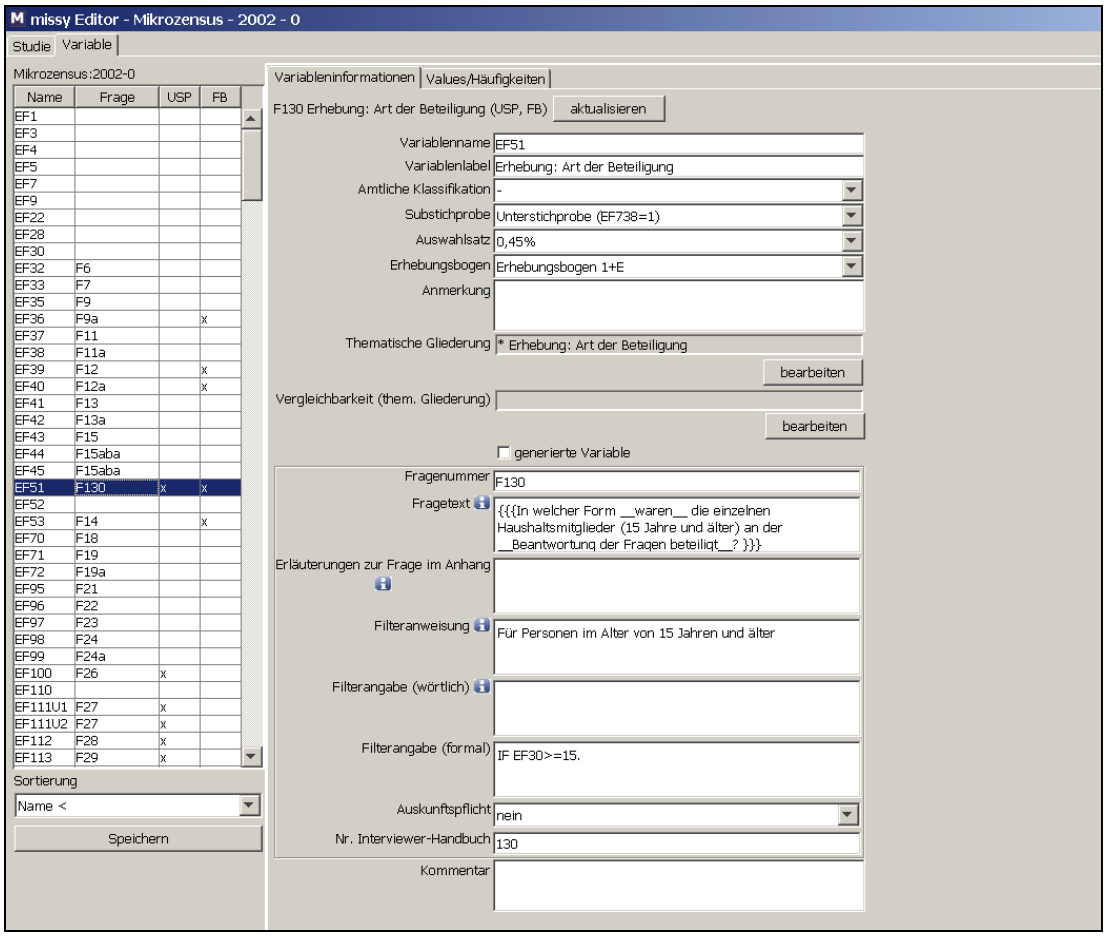
Auf dieser Seite werden alle Informationen erfasst, die in MISSY II auf der Ebene der Detailinformationen zu sehen sind (vgl. Abb. 1). D. h. für alle Merkmale des enthaltenen Datensatzes werden hier der Name und das Label der Variable sowie der zugehörige Erhebungsbogen erfasst. Zudem gibt es weitere Felder, in denen der jeweilige Auswahlsatz einer Variable, eine ggf. zugrundeliegende amtliche Klassifikation, die Kennzeichnung als Substichprobenmerkmal und die Einordnung in die thematische Gliederung angegeben werden können. Variablennamen und -label, Values, Valuelabel sowie die Häufigkeiten werden nicht von Hand eingegeben, sondern aus dem XML Output eines Statistiksystems nach einer Konvertierung in DDI 3.0 direkt in das System eingebunden (vgl. Kapitel 2.1). Die Variablen- und Valuelabels sind über Freitextfelder erfasst, so dass gegebenenfalls eine Überarbeitung möglich ist. Alle

anderen genannten Felder enthalten vollständige Listen, aus denen der entsprechende Eintrag ausgewählt werden kann.

Bei der weiteren Eingabe ist entscheidend, ob es sich bei dem betreffenden Merkmal um eine generierte oder eine erhobene Variable handelt. Für „generierte Variablen" erscheinen im Editor andere Auswahlmöglichkeiten, da hier bestimmte Angaben z. B. zum Fragetext entfallen. Relevant ist für generierte Merkmale auf Detailebene dagegen, ob es sich um eine Bandsatzerweiterung oder eine Typisierung handelt, auf welche Einheit sich die Angabe bezieht und ob der Variablen ggf. ein bestimmtes amtliches Konzept zugrunde liegt.

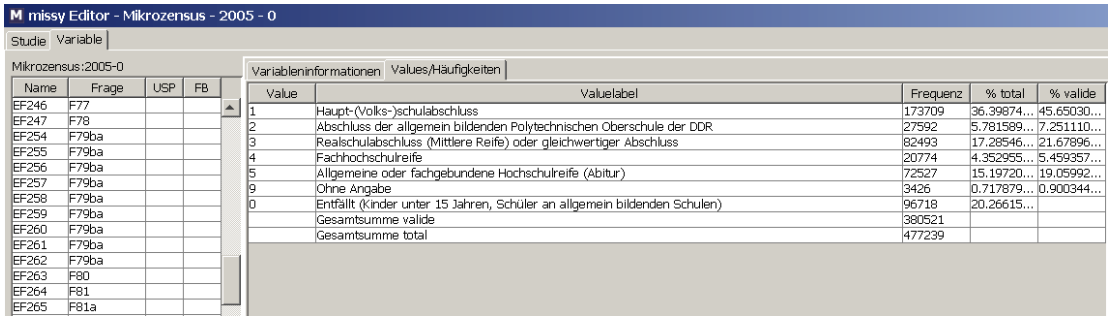
Für direkt erhobene Merkmale müssen dagegen u. a. Angaben zu der Fragenummer, dem Fragetext, der gesetzlichen Auskunftspflicht und der Filterangabe gemacht werden.

Abbildung 1: Eingabemaske der Detailinformationen des MISSY-Editors



Über die Karteikarte „Values/Häufigkeiten" auf der Variablenseite des Editors gelangt man zu der Häufigkeitsauszählung der ausgewählten Variable (vgl. Abb. 2). Diese Eingabemaske dient der Kontrolle und gegebenenfalls der Korrektur bestehender Value Labels.

Abbildung 2: Ausschnitt der Bearbeitungsmaske zu den Value Labels und Häufigkeiten des MISSY-Editors



Während der gesamten Arbeit an den Dateien befinden sich diese auf einem zentralen Server, der direkt mit dem Informationssystem verbunden ist. Durch eine Überwachung der Dateien durch das Informationssystem werden die Daten regelmäßig automatisch aktualisiert. Der Stand aller bereits eingegebenen Informationen wird täglich durch eine Backupsoftware gesichert und zusätzlich täglich in ein Versionskontrollsystem eingecheckt, um Datenverluste aufgrund von Problemen mit dem Server oder netzbedingtem Datenverlust etc. auszuschließen.

1.2 Aufbereitung und Eingabe der Metadaten

Ein Schwerpunkt der inhaltlichen Arbeiten bestand in der Aufbereitung und anschließenden Eingabe der variablenbezogenen Informationen zu den verfügbaren Mikrozensus Scientific Use Files von 1973 bis 2006. Die Informationen zu den Mikrozensen 1995 und 1997 wurden schon im Rahmen des Pilotprojekts in das System eingebunden, so dass sich die Arbeit im ersten Jahr der zweiten Projektphase auf die insgesamt 11 Files der Erhebungsjahre 1989, 1991, 1993, 1996 sowie 1998-2004, im zweiten Jahr auf die insgesamt neun Datensätze der Erhebungsjahre 1973, 1976, 1978, 1980, 1982, 1985 und 1987, sowie 2005 und 2006 bezog. In MISSY werden für jede Variable im Datensatz alle zugänglichen Informationen auf einer Seite zusammengefasst (sog. Variableninformationen). Dabei können nur wenige der Informationen automatisch generiert werden. Meta-Informationen wie Variablennamen, Values und Valuelabels können zwar über spezielle Tools direkt eingelesen werden, allerdings geht der Inhalt von MISSY weit über die Vermittlung dieser Basisinformationen hinaus. Die Zielsetzung von MISSY besteht vielmehr darin, für jede Variable möglichst umfassende Informationen bereit zu stellen und diese sind gegenwärtig noch nicht an einer Stelle konzentriert verfügbar und einfach über Skript einlesbar, sondern müssen aus einer Reihe verschiedener Dokumentationen zusammengetragen, aufbereitet und in strukturierter Form über das Erfassungswerkzeug eingegeben werden. Im Einzelnen handelt es sich um die folgenden Informationen:

- Die Variablenlabels wurden in der ersten Projektphase von MISSY für die Variablen der Mikrozensen 1995 und 1997 sprachlich überarbeitet und vereinheitlicht und folgen einem festgelegten Schema. Diese sprachliche Überarbeitung war notwendig, da nur mit einheitlichen Variablenlabels eine intertemporale Vergleichbarkeit der Variablen hergestellt werden kann. Für die Einbindung der weiteren Scientific Use Files in MISSY musste insbesondere die Anpassung und Überarbeitung der Variablenlabels aus den Ergänzungs- und Zusatzprogrammen geleistet werden.

- Da sich das Problem der mangelnden „Datenbanktauglichkeit" auch für die aus den SPSS-Setups generierten Valuelabels stellt, wurden diese manuell nachgebessert. Diese Nachbearbeitung umfasste vor allem die Überprüfung der Rechtschreibung sowie die Vereinheitlichung der Schreibweise (übliche Abkürzungen etc.). Besonders in den älteren Setups war dies mit einem hohen Arbeitsaufwand verbunden, da ältere SPSS-Versionen sehr restriktiv in der zulässigen Länge der Value-Labels waren, welches besonders bei komplexen Sachverhalten (z. B. Familientypisierungen) starke Kürzungen der Value Labels zur Folge hatte, deren Verständlichkeit ohne Hinzuziehen weiterer Informationen (z. B. Fragebogen oder Schlüsselverzeichnis) kaum möglich war. Hier war entsprechend eine sehr sorgfältige Überarbeitung nötig.
- Die Informationen zu einer Frage im Fragebogen umfassen den Namen des Fragebogens, Fragenummer und -text sowie Erläuterungen zur Frage im Anhang des Fragebogens. Bei der Eingabe dieser Informationen wurden nicht nur die genauen Frageformulierungen, sondern auch bestimmte Formatierungen wie z. B. Hervorhebungen aus dem Fragebogen übernommen.
- Für Variablen, die auf Fragen basieren, die nur einer Auswahl von Befragten gestellt werden, wurden – soweit möglich – die Filteranweisungen und Filterführungen des Fragebogens nachvollzogen. Diese Darstellungen wurden für jedes Jahr in der Regel anhand des Fragebogens erarbeitet und sowohl formal (in SPSS-Syntax) als auch sprachlich umgesetzt. Da bei einigen Mikrozensen (1973, 1976, 1978, 1980) nur noch die Interviewerbögen verfügbar sind und die Filterführung in den älteren Mikrozensen ohnehin nicht immer eindeutig ist, mussten bestehende Filter z. T. aus den Interviewerhandbüchern, aber auch aus den Daten selbst, erschlossen werden.
- Für die Variablen, die nicht zum Grundprogramm des Mikrozensus gehören, wurde dokumentiert, zu welchem Ergänzungs- oder Zusatzprogramm sie gehören (Substichprobe) und mit welchem Auswahlsatz dieses Programm erhoben wurde.
- Für jede Variable wurde dokumentiert, ob die zugrunde liegende Frage für alle oder einen Teil der Befragten auf freiwilliger Basis beantwortet wird (Auskunftspflicht).
- Bei den Variablen, die auf Basis anderer Merkmale generiert wurden, wurden Informationen zur Art der Variable, zur Einheit, auf die sie sich bezieht, sowie zu dem ihnen zugrundeliegenden Konzept erfasst. Diese Informationen werden in MISSY II verwendet, um die Erstellung und Sortierung verschiedener Listen und Zugänge zu generierten Variablen zu ermöglichen.

Da sich das Erhebungskonzept des Mikrozensus ab 2005 in wichtigen Punkten geändert hat, mussten diese Modifikationen konzeptuell in MISSY berücksichtigt werden. Neben einer Erweiterung und erheblichen Veränderungen des Erhebungsprogramms, das insbesondere auf Ebene der Korrespondenzmatrix berücksichtigt werden musste (vgl. Punkt 1.3), ist vor allem das Konzept der Unterjährigkeit zu erwähnen, für das eine Dokumentation auf der Hintergrundebene erarbeitet wurde. Vor diesem Hintergrund wurde die Studienbeschreibung überarbeitet und um die Neuerungen ab 2005 erweitert. Bei dem Konzept, das hierfür entwickelt und umgesetzt wurde, handelt es sich um eine für alle verfügbaren Mikrozensus Scientific Use Files allgemeingültige Studienbeschreibung, von der aus ein schneller Zugriff auf die Besonderheiten und Veränderungen in den einzelnen Jahren bzw. Erhebungszeiträumen möglich ist, und in die zukünftige Veränderungen problemlos integriert werden können. Dies hat den Vorteil, dass Nutzer alle Informationen an einer zentralen Stelle finden und die Studienbeschreibung bei der Einbindung eines neuen Mikrozensusjahres nicht komplett überarbeitet, sondern nur an den relevanten Stellen aktualisiert werden muss.

Aufbereitung von PDF Dokumenten

In der Projektlaufzeit wurden für die verfügbaren Mikrozensus-Jahrgänge von 1973-2006 die erforderlichen Hintergrundinformationen zu den Variablen zusammengetragen und für die Einpflege ins Datenbanksystem bearbeitet. Dies umfasste Interviewerhandbücher, Fragebögen sowie Schlüsselverzeichnisse der einzelnen Mikrozensus-Jahrgänge. Die Beschaffung und Bearbeitung der Dokumente war mit unterschiedlich hohem Aufwand verbunden. Einige Interviewerhandbücher standen nicht zur Verfügung und wurden beim Statistischen Bundesamt angefordert. Für die älteren Mikrozensen sind die Fragebögen für Selbstausfüller nicht mehr existent, sodass hier auf die Erhebungslisten für die Interviewer zurückgegriffen werden musste. Mehrere Dokumente lagen nicht in elektronischer Form vor und wurden zunächst manuell eingescannt. Dies betraf die Interviewerhandbücher 1976-1982, 1985, 1987, 1989, 1991, 1993 und 1996 sowie die Interviewerlisten der Jahre 1973, 1976, 1978, 1980 und 1982. Da die vorliegenden Dokumente von sehr unterschiedlicher Qualität waren, wurden die eingescannten Dokumente im Anschluss für den Web-Zugriff optimiert. Alle Dokumente wurden im PDF-Format gespeichert und mit Sprungmarken (sog. „Named Destinations") versehen, wobei im Berichtszeitraum jede einzelne Frage-, Variablen- und Interviewerhandbuchnummer der Mikrozensen 1973-2006 als eine solche Sprungmarke versehen werden musste.[2] Sprungmarken ermöglichen die gezielte Verlinkung aus den Detailinformationen der Variablen auf die entsprechenden Stellen in den PDF-Dokumenten, sodass Nutzer bei Interesse an den Originaldokumenten gleich die betreffende Stelle zum interessierenden Merkmal finden.

1.3 Erweiterung der Variablen-Zeitpunkte-Matrix

Die Matrix erfüllt mehrere Funktionen gleichzeitig und stellt damit ein Kernstück von MISSY dar: Zunächst gibt die Matrix eine thematisch gegliederte Übersicht über alle in den Mikrozensus SUF enthaltenen Merkmale. Auf einen Blick werden dabei die Vergleichbarkeit von Merkmalen über die Zeit sowie sog. Variablenbrüche zwischen Erhebungsjahren oder Erhebungszeiträumen ersichtlich. Darüber hinaus ermöglicht die Matrix einen schnellen und direkten Zugang zu den variablenbezogenen Informationen der interessierenden Merkmale. Die Matrix bedient damit Ansprüche ganz unterschiedlicher Mikrozensus-Nutzer: Personen, die nicht mit dem Datensatz vertraut sind, können sich über die vollständige Matrix (vgl. Abb. 3) schnell einen Überblick verschaffen, welche Themenfelder im Mikrozensus abgedeckt werden, für welche Zeitpunkte diese vorliegen und wie tief gegliedert die Angaben im Einzelfall sind. Erfahrene Mikrozensus-Nutzer werden vor allem von der Vergleichbarkeit über die Erhebungsjahre profitieren, wobei die Option der individuellen Zusammenstellung interessierender Erhebungsjahre und Themenfelder sie in besonderer Weise unterstützt.

[2] Bei den Erhebungslisten wurde auf Sprungmarken verzichtet, da die Dokumente selten mehr als drei Seiten umfassen.

Abbildung 3: Ausschnitt der vollständigen Matrix

Thematische Gliederung	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995
• Demographie und Bevölkerung												
•• Daten zur Person												
••• Alter	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995
Alter	EF44	EF44	EF30	EF30	EF30	EF30	EF30	EF30	EF30	EF30	EF30	EF23
Alter: Haushaltsbezugsp.	EF754	EF754	EF558	EF558	EF558	EF558	EF558	EF558	EF558	EF558	EF558	EF188
Alter: Familienbezugsp.			EF593	EF593	EF593	EF593	EF593	EF593	EF593	EF593	EF593	EF211
Alter: Bezugsp. der Lebensform	EF820	EF820										
Alter: Ehefrau der Familienbezugsp.			EF611	EF611	EF611	EF611	EF611	EF611	EF611	EF611	EF611	EF223
Alter: Lebenspartner der Haushaltsbezugsp.	EF844	EF844	EF659	EF659	EF659	EF659	EF659	EF659	EF659	EF659	EF659	
Alter: Lebenspartner der Bezugsp. der Lebensform	EF844	EF844	EF659	EF659	EF659	EF659	EF659	EF659	EF659	EF659	EF659	
Alter: Haupteinkommensbezieher	EF732	EF732										
Alter: Ernährer	EF732	EF732										
Geburtsjahr	EF47	EF47	EF33	EF33	EF33	EF33	EF33	EF33	EF33	EF33	EF33	EF37
Geburtsmonat												
••• Geschlecht	2006	2005	2004	2003	2002	2001	2000	1999	1998	1997	1996	1995
Geschlecht	EF46	EF46	EF32	EF32	EF32	EF32	EF32	EF32	EF32	EF32	EF32	EF35
Geschlecht: Haushaltsbezugsp.	EF753	EF753	EF557	EF557	EF557	EF557	EF557	EF557	EF557	EF557	EF557	EF187
Geschlecht: Familienbezugsp.			EF592	EF592	EF592	EF592	EF592	EF592	EF592	EF592	EF592	EF210
Geschlecht: Bezugsp. der Lebensform	EF819	EF819										
Geschlecht: Lebenspartner der	EF843	EF843	EF657	EF657	EF657	EF657	EF657	EF657	EF657	EF657	EF657	

Bei der individuell zu erzeugenden Matrix können Nutzer aus der thematischen Gliederung einzelne Themenfelder sowie Erhebungszeitpunkte auswählen (vgl. Abb. 4). Damit wird die Matrix auf die interessierenden Variablen der jeweiligen Mikrozensen eingeschränkt. Auf diese Weise wird sichergestellt, dass die Matrix mit der Fülle an enthaltenen Informationen, welche zunächst ein großer Verdienst von MISSY ist, zugleich gezielt suchende Nutzer nicht mit (im Moment) unerwünschten Angaben überfrachtet. Somit wird bei eingeschränkter Suche der Zugang zu bestimmten Informationen der Matrix auch kognitiv erleichtert.

In der Matrix sind bestimmte Merkmale farblich unterlegt. Mit dieser Kennzeichnung wird auf einen Blick ersichtlich, dass der Vergleich über die Zeit innerhalb der betreffenden Zeile nicht trivial ist. So kann die Markierung zum Beispiel darauf verweisen, dass der Inhalt eines Merkmals in einem anderen Jahr in zwei Variablen erfasst wird. In diesem Fall werden beide Variablen farblich unterlegt als Vergleichsvariablen angezeigt. Der Vergleich kann aber auch dadurch erschwert werden, dass sich der Inhalt der Variable in einem Maße geändert hat, dass keine direkte Vergleichbarkeit – wie sie sonst über die Zeilen angezeigt wird – gegeben ist. Das ist beispielsweise dann der Fall, wenn das zugrunde liegende Konzept eines Merkmals modifiziert worden ist (so kommt z. B. anstelle des Konzeptes Familie ab 2005 nur noch das Konzept der Lebensform zur Anwendung; vgl. die unterlegten Zeilen in Abb. 3).

Die Erstellung der Matrix hat sich über verschiedene Projektphasen von MISSY erstreckt: In der ersten Projektphase wurde eine Matrix mit allen Merkmalen der verfügbaren Mikrozensus SUF von 1989 bis 2004 erstellt. Die Informationen zur Vergleichbarkeit wurden optimiert. Variablen, die in der Matrix in einer Zeile (ohne farbliche Markierung) stehen, sind zumindest „auf der obersten Ebene" vergleichbar. Da der Mikrozensus, in dem vorliegenden – bereits über 30 Jahre umspannenden – Zeitraum, deutlich weiterentwickelt wurde und Angaben immer detaillierter abgefragt und den veränderten rechtlichen

Rahmenbedingungen angepasst wurden, wäre eine zu restriktive Anforderung an das Kriterium der Vergleichbarkeit kontraproduktiv gewesen. Um dennoch auf kleinere Unterschiede zwischen vergleichbaren Variablen verschiedener Erhebungsjahre hinzuweisen, werden spezifische Zusätze in den Variablen Labels auf Ebene der variablenbezogenen Informationen angezeigt (z. B. auf welchen Monat sich die Einkommensangabe im jeweiligen Jahr bezieht).

In der zweiten und dritten Projektphase wurde die Matrix um die Erhebungsjahre 2005 und 2006 sowie um die Erhebungsjahre vor 1989 erweitert. Aufgrund der in diesen Jahren neu hinzugekommenen Variablen bzw. Themenkomplexe wurde die Matrix deutlich ausgeweitet. Einerseits wurde die Matrix dabei inhaltlich um zahlreiche Zeilen ergänzt, welche in die bestehende thematische Gliederung eingepasst oder – wo dies nicht möglich war – in neu erstellte Gliederungspunkte aufgenommen wurden. Sämtliche Merkmale der hinzugefügten Mikrozensus SUF mussten dabei auf Vergleichbarkeit mit den bereits in der Matrix enthaltenen Merkmalen geprüft werden. Variablenbrüche wurden – wie bisher – mit einer farblichen Markierung gekennzeichnet.

Abbildung 4: Auswahl der individuellen Matrix

1.4 Ausbau des inhaltlichen Service

Im Pilotprojekt wurde damit begonnen, die Informationen zu den Daten um einen inhaltlichen Service zu ergänzen. Dieser Service wurde in MISSY II ausgeweitet, mit dem Ziel, den Aspekt der Wissensvermittlung online umzusetzen. Die Nutzer sollen dazu angeregt und dabei unterstützt werden, sich inhaltlich mit den Daten auseinanderzusetzen.

Zum einen soll der inhaltliche Service das Analysepotential des Mikrozensus aufzeigen. Potentielle Nutzer, die bisher noch nicht mit den Daten des Mikrozensus gearbeitet haben, können sich so einen Überblick über die mit dem Mikrozensus bearbeitbaren Fragestellungen verschaffen und prüfen, ob die

Daten für ihre eigene Fragestellung geeignet sind. Auch Anregungen für mögliche Forschungsfragen werden geliefert.

Neuen Nutzern des Mikrozensus soll der Einstieg in das Arbeiten mit den Daten erleichtert werden. Dazu werden grundlegende Informationen bereitgestellt, die für erste Aufbereitungen und Auswertungen der Daten unerlässlich sind. Hierzu gehören auch Hinweise auf Besonderheiten des Mikrozensus, die in anderen Datenquellen so nicht vorhanden sind und deshalb leicht übersehen werden. Für eine adäquate Analyse der Daten ist die Beachtung dieser Besonderheiten grundlegend.

Für erfahrene Nutzer des Mikrozensus sollen komplexe Werkzeuge zur Auswertung der Daten bereitgestellt werden. Sie dienen zum einen der Umsetzung sozialwissenschaftlicher Konzepte mit den Daten des Mikrozensus, so dass die Anwendung dieser Konzepte ermöglicht und zugleich vereinheitlicht wird. Zum anderen werden die in der amtlichen Statistik verwendeten Konzepte erläutert und so der sozialwissenschaftlichen Analyse zugänglich gemacht.

Mit dem inhaltlichen Service des Datenbanksystems MISSY II sind diese Ziele folgendermaßen umgesetzt:

- Um das Analysepotential des Mikrozensus aufzuzeigen, sind die bereits im Pilotprojekt zusammengestellten inhaltlichen Tabellen in das nun vollständig ausgebaute Informationssystem eingebunden (Arbeitshilfen – Datenanalyse – Inhaltliche Tabellen). Somit stehen zu den Schwerpunktthemen des Mikrozensus, wozu vor allem Erwerbstätigkeit, Bildung und Einkommen sowie Haushalts-, Familien- und Lebensformen zählen, beispielhafte Auswertungen zur Verfügung. Zur Veranschaulichung sind einige Ergebnisse dieser Auswertungen auch grafisch dokumentiert.
- Anhand von Beispielen wird gezeigt, wie einfache Auswertungen des Mikrozensus konkret umgesetzt werden (Arbeitshilfen – Datenanalyse – Auswertungsbeispiele). Die zur Beantwortung einer Fragestellung notwendigen Schritte der Aufbereitung und Auswertung der Daten werden einzeln erläutert und die dazu notwendige Syntax wird – sowohl in SPSS als auch in Stata – zur Verfügung gestellt. Die Beispiele sind so konzipiert, dass sie auch von unerfahrenen Nutzern eigenständig nachvollzogen werden können.
- Neben verschiedenen inhaltlichen Auswertungen sind dabei auch typische Probleme im Umgang mit den Daten aufgegriffen. So wird beispielsweise gezeigt, worin sich Auswertungen auf Personen- und Haushaltsebene unterscheiden und wie sich die jeweils relevante Einheit festlegen lässt. Aus einem anderen Beispiel geht hervor, wie sich eine eindeutige und fortlaufende Haushaltsnummer generieren lässt.
- Die wichtigsten Schritte der Datenaufbereitung sind zusätzlich separat zusammengestellt (Arbeitshilfen – Datenaufbereitung). Hier erfährt der Nutzer, welche Entscheidungen er vor Beginn einer Auswertung treffen muss und welche Besonderheiten des Mikrozensus es zu beachten gilt. Um die einzelnen Schritte schnell und einfach umsetzen zu können, sind Links zu den jeweils benötigten Variablen gesetzt. So hat der Nutzer für alle Erhebungsjahre des Mikrozensus beispielsweise sofort im Blick, über welche Variablen und über welche Ausprägungen dieser Variablen er die Bevölkerung in Privathaushalten abgrenzen kann.
- Die im Service-Angebot des GML bereits vorhandenen Mikrodaten-Tools, die den praktischen Umgang mit den Daten in unterschiedlichen Themenbereichen dokumentieren, sind in MISSY II in-

tegriert (Arbeitshilfen – Mikrodatentools). Die Nutzer haben Zugang zu verschiedenen Arbeitspapieren, in denen die Mikrodaten-Tools eingehend erläutert sind, und können die zu deren Umsetzung notwendige Syntax (in SPSS und Stata) herunterladen.

Angelegt ist der inhaltliche Service in MISSY II so, dass er systematisch erweitert werden kann. Weitere Informationen zur Datenaufbereitung und -analyse sowie zusätzliche Auswertungsbeispiele, die beispielsweise in Workshops, bei der Beratung oder der Forschungstätigkeit des GML entstehen, können problemlos eingefügt werden.

2 Informationstechnologische Arbeiten

2.1 DDI 3.0 Editor

Basierend auf der Architektur des Fragebogeneditors aus dem Projekt QDDS wurde ein Editor entwickelt, der sich an den Daten orientiert, die für die Darstellung der Dokumentation des Mikrozensus notwendig sind. Aus dieser Vorgehensweise ergaben sich verschiedene Vorteile: Erstens musste keine neue Architektur entwickelt werden, zweitens waren zentrale Komponenten wieder verwendbar und drittens ergibt sich eine Durchlässigkeit zwischen den beiden Editoren, die für zukünftige Entwicklungen genutzt werden kann.

Der MISSY Editor arbeitet grundsätzlich netzgestützt. Konkret bedeutet das, dass er die Datei eines Erhebungszeitraumes von einem zentralen Server lädt, diese dort für weitere Ladevorgänge sperrt, die Datei spätestens zum Ende der Editierarbeiten speichert und schließlich die Sperre wieder löst. Somit bleibt gewährleistet, dass nicht mehrere Personen gleichzeitig an einem Erhebungsjahr arbeiten und sich gegenseitig Änderungen überschreiben könnten. Ein gesperrter Erhebungszeitraum bleibt allerdings für einen rein lesenden Zugriff durch den Editor zugänglich. Das Interface des Editors informiert den Nutzer dann, dass die Datei schreibgeschützt geöffnet wurde.

Im Wesentlichen nutzt der Editor für seine Speicherung den Datenstandard DDI 3.0. Lediglich solche Informationen, für die es im Datenstandard keine direkte Abbildung gibt, wurden mittels einer konsistenten Notation zur Benennung in Kommentarknoten des Formats abgelegt. Eine Ausnahme bildet ein Eingabefeld, das zur internen Kommunikation zwischen den Nutzern des Editors gedacht ist. Dessen Inhalte werden in einer getrennten Struktur direkt auf dem Server gespeichert.

Neue Erhebungszeiträume des Mikrozensus werden im aktuellen Stand so angelegt, dass der XML Output eines Statistiksystems in eine DDI 3.0 Datei konvertiert und auf dem Server abgelegt wird. Diese initiale Version eines Erhebungszeitraums enthält bereits sämtliche Variablen mit ihren Namen, Label, ihren Values und Valuelabel (incl. Häufigkeiten) und die Nummern der Fragen. Perspektivisch soll dieser Schritt nicht mehr durch einen Administrator abgewickelt werden, sondern durch den Nutzer selbst, indem dieser direkt die Ausgabedatei des Statistiksystems im Editor auswählt, um einen Erhebungszeitpunkt hinzuzufügen.

Um den Editor für andere Studien verwendbar zu machen sind zwei Punkte ausschlaggebend: Erstens muss das Feldschema für diese Studie passend sein und zweitens müssten die verwendeten Vokabulare für diese Studie definiert werden. Strukturell ist der Editor darauf bereits vorbereitet.

2.2 DDI-Service / Metadatenbank

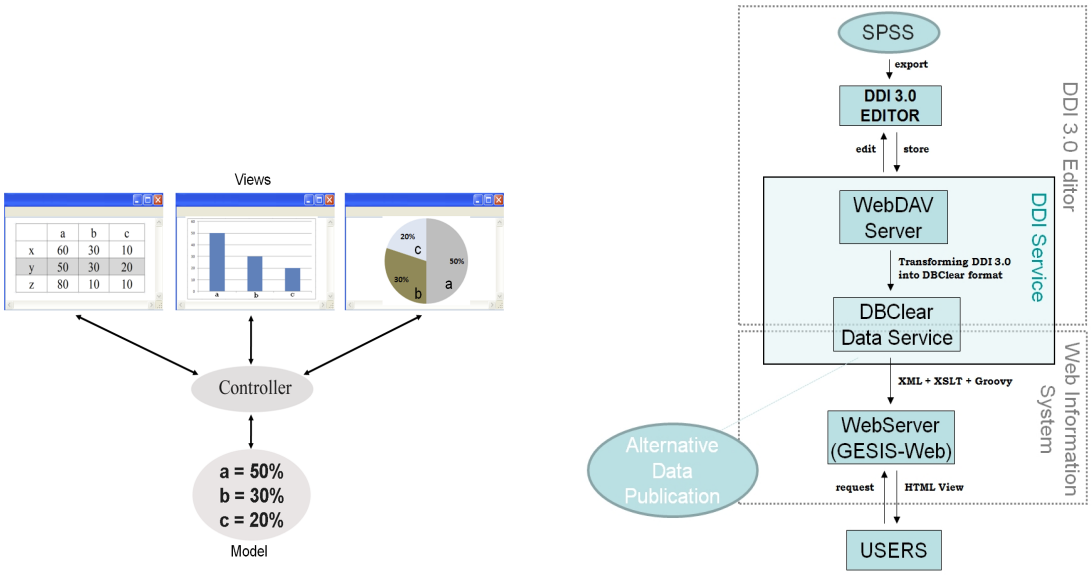
Der DDI 3.0-Editor erzeugt Textdateien, wobei jede Textdatei ein Erhebungsjahr darstellt und tausende von Zeilen enthält. In einer Textdatei werden die Metadaten eines Erhebungsjahres als ein einzelner Datensatz betrachtet, obwohl sie bereits hierarchisch mit DDI 3.0 Standard in XML strukturiert und geschrieben wurden. Wenn in den Textdateien nach einer bestimmen Variablen eines bestimmtes Erhebungsjahres gesucht werden soll, führt das natürlich zu einer sehr langen Rechenzeit. Um einen Klartext, z. B. Fragetext, zu extrahieren, muss jede Zeile in den Dateien ausgelesen werden; dafür scheint die Rechenzeit zu hoch. Daher wurde entschieden, die DDI 3.0-Dateien in eine so genannte Metadatenbank zu überführen[3], da die Umwandlung der DDI 3.0 in die Datenbank die Rechenzeit reduziert, z. B. bei der Suche oder bei der Informationsdarstellung.

Als Metadatenbank wird DBClear verwendet. DBClear ist ein allgemeines, plattform-unabhängiges Clearinghouse-System, dessen Metadaten-Schema auch an andere Standards angepasst werden kann. DBClear hat eine offene Architektur und wiederverwendbare Komponenten, die für andere Anwendungen relativ leicht anzupassen sind. Darüber hinaus bietet DBClear auch Möglichkeiten zum Aufbau eines Web-Informationssystems, das mit dem MVC (Model-View-Controller) Design Pattern übereinstimmt.

Ein MVC Design Pattern wird in der Regel dazu genutzt, um eine schnelle und effektive Lösung für häufig auftretende Probleme in der Software-Entwicklung zu finden und ist damit auch für die Entwicklung von Web-basierten Software-Anwendungen geeignet. In Abbildung 5(a) sind die Struktur und die Wirkungsweise des MVA dargestellt. Dieses Modell wurde für die gespeicherten Metadaten in der Metadatenbank angewandt und aus dem MVC Design Pattern wurde die Architektur der Anwendungs-Software abgeleitet (vgl. Abbildung 5(b)). Die Abbildung verdeutlicht, welche Komponente welche Aufgabe in der Software übernimmt und wie sich die einzelnen Teile der Software voneinander unterscheiden.

[3] Sofern nicht anders angegeben, ist eine Metadatenbank eine Datenbank für Metadaten.

Abbildung 5: Anwendungssoftware



(a) Model-View-Controller Design Pattern (b) Software-Architektur

Die DDI 3.0-Umwandlung in das DBClear-Metadatenbank-Format ist eigentlich ein „Verflachungs"-Prozess von einer hierarchischen Struktur auf eine tabellarische Struktur. Eine detaillierte Erklärung der hierarchischen Struktur von DDI 3.0 findet man auf der Website der DDI-Allianz.[4] Abbildung 6 zeigt einen kurzen Überblick über die DDI 3.0-Struktur. Von diesem Aufbau ausgehend wird die DDI 3.0-Struktur in eine tabellarische Struktur umgewandelt (vgl. Tabelle 1), wo jeder Datensatz als eine Ressource gilt. Das Format ist prinzipiell mit einem zweidimensionalen Daten-Modell vergleichbar, wie in Tabelle 1 dargestellt. Diese tabellarische Struktur wird in eine XML-Spezifikation überführt (vgl. Abb. 6), deren Metadaten-Schema als DBClear-Schema bezeichnet wird. Die DDI 3.0-Struktur wird in das DBClear-Schema unter Verwendung der XSL-Transformation[5] konvertiert. Es wird durch DBClear in ein eigenes RDBMS-Schema (Relational Database Management System) übersetzt. DBClear ist mit mehreren RDBMS-Softwares kompatibel (z. B. Oracle, MySQL, PostgreSQL); in diesem Fall wird PostgreSQL[6] benutzt.

[4] http://www.ddialliance.org/
[5] XSL steht für Extensible Stylesheet Language. http://www.w3schools.com/xsl/
[6] PostgreSQL ist eine Open-Source-RDBMS-Software. http://www.postgresql.org/

Abbildung 6: DDI 3.0-Hierarchie

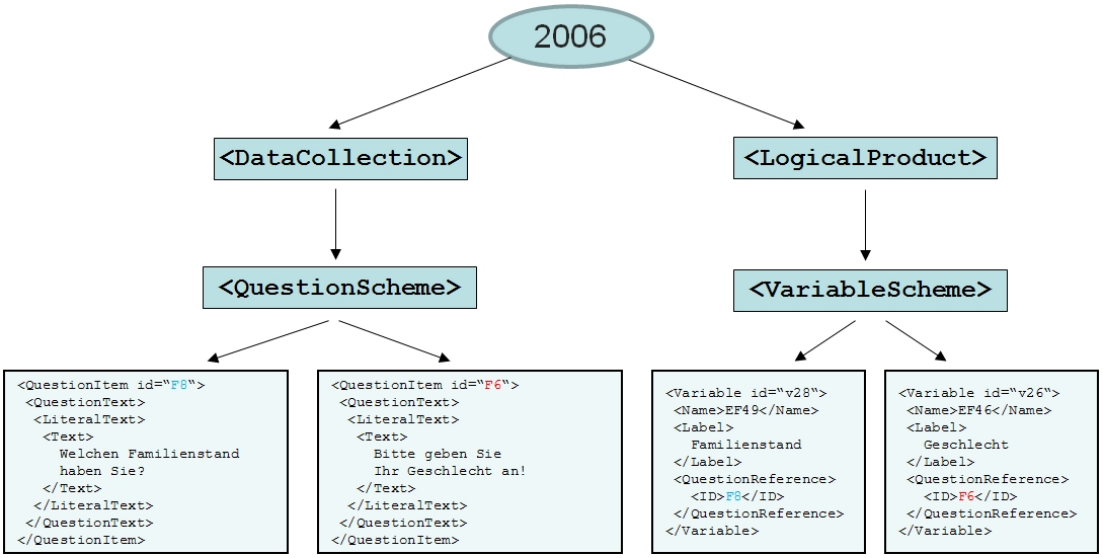


Tabelle 1: Tabellarische Struktur als grundlegendes Schema des DBClear-Formats

Variable	*Erhebungsjahr*	*Fragenummer*	*Fragetext*
EF46	2006	F6	Bitte geben Sie Ihr **Geschlecht** an!
EF49	2006	F8	Welchen **Familienstand** haben Sie?
...	...	...	...

2.3 Web-Informationssystem

Neben der DDI 3.0-Umwandlung in das DBClear-Metadatenbank-Format bestand ein weiterer wichtiger Aspekt der informationstechnologischen Implementierung darin, den Nutzerzugang zu den Metadaten in MISSY zu verbessern. Hierzu wird das Softwaresystem DBClear eingesetzt. DBClear enthält als Alleinstellungsmerkmal eine Browsing-Komponente, mit der die Nutzer mittels hierarchischer Inhaltsverzeichnisse auf die gespeicherten Informationen zugreifen können (Faceted Browsing). Diese Inhaltsverzeichnisse werden auf der Basis von Metadaten dynamisch generiert, d. h. es müssen nicht alle gewünschten Blätterstrukturen im System hinterlegt werden, sondern werden vom System bei Bedarf mittels Filterung generiert. Der Nutzen des Faceted Browsing wurde bereits in verschiedenen Anwendungen bestätigt.[7]

[7] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. ACM SIGCHI: Human Factors in Computing Systems, 2003.

Auf der Basis des früheren Projekts werden den Endnutzern drei Blätterzugänge angeboten: (a) die „Variablenliste", (b) die „Thematische Gliederung" und (c) die „Variablen-Zeitpunkte-Matrix". Weitere Blätterzugänge sind prinzipiell möglich. Die DBClear-Anwendung erzeugt für jede Anfrage ein XML-Dokument. Dieses XML-Dokument ist nur maschinell lesbar und muss daher in ein gültiges HTML-Dokument umgewandelt werden. Durch den Einsatz des Browsers wird das Dokument für die Endnutzer lesbar. Um XML in HTML-Dokumente zu transformieren, werden wiederum XSL-Transformationen verwendet, da diese es erlauben, Umwandlungen gemäß den jeweiligen Anforderungen vorzunehmen, ohne die Datenquelle zu ändern. Darüber hinaus wird die Umwandlung mit Hilfe von Java und Groovy[8], eingebettet in XSL, verbessert. Da MISSY in das Online-Angebot der GESIS integriert ist, wird für die Webanwendung von MISSY das für das GESIS-Web etablierte Content-Management-System Typo3[9] genutzt.

Abbildungen 7 bis 9 geben einen Überblick über einige wichtige Funktionen der Webanwendung. In Abbildung 7 ist die Detailansicht einer Variablen dargestellt, in der alle wichtigen Informationen (vgl. Kapitel II.1.2) zu der ausgewählten Variablen aufgeführt sind. In den Detailansichten ist auch jeweils ein Überblick über die Variable zu anderen Erhebungsjahren enthalten, wodurch die zeitliche Vergleichbarkeit deutlich erleichtert wird. Über so genannte „Tooltips" lassen sich zusätzliche Erläuterungen zu den einzelnen Informationsfeldern abrufen.

Abbildung 7: Detailansicht einer Variablen

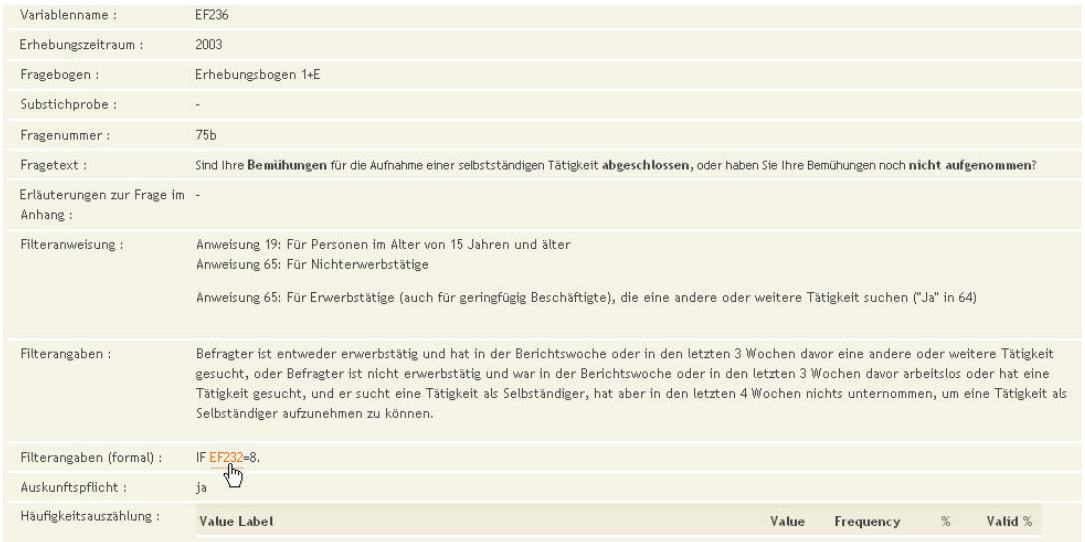
[8] Groovy ist eine dynamische Sprache für die Java-Plattform. http://groovy.codehaus.org/
[9] http://www.typo3.com/

Abbildung 8 verdeutlicht den Nutzerzugang über die thematische Gliederung. Dieser Zugang erlaubt es dem Nutzer, die Suche unter einem bestimmten Themengebiet zu spezifizieren. Aufgrund der hierarchischen Struktur der thematischen Gliederung gelangt der Nutzer auf der letzten Stufe zu einer Auflistung der zu dem ausgewählten Themenbereich verfügbaren Variablen. Nach Auswahl des interessierenden Erhebungsjahres wird wiederum die Detailansicht der Variable angezeigt.

Abbildung 8: Nutzerzugang Thematische Gliederung

Da die Variablennamen und Erhebungsjahre als kontrollierte Vokabulare erfasst sind, können sie aufgrund ihrer eindeutigen Kennung in der Metadatenbank verwendet werden, um Hyperlinks innerhalb der Informationsfelder zu generieren. In Abbildung 9 wird dieses Feature am Beispiel der Verlinkung einer Filtervariable aufgezeigt.

Abbildung 9: Aus der Datenbank generierte Hyperlinks



3 Nutzen und Verwertbarkeit

Das Ziel des MISSY-Projekts besteht im Aufbau eines umfassenden Informationssystems für Mikrodaten, das die bisherigen Service-Seiten zur Datendokumentation des German Microdata Labs ablösen soll. Im Vordergrund steht dabei die Bereitstellung strukturierter Metadaten nach einem einheitlichen Standard und die Optimierung der Aufbereitungsprozesse der zu dokumentierenden Datensätze. Durch das im Projekt entwickelte Erfassungswerkzeug für DDI-konforme Metadaten wird das Einpflegen von Metadaten effizienter und ist zudem offen für Anpassungen an künftige Entwicklungen. Als eine mögliche Erweiterung von MISSY ist beispielsweise der Aufbau eines englischen Metadatenangebots zu europäischen Daten zu nennen.

4 Fortschritt bei anderen Stellen

Neben MISSY sind derzeit drei Systeme online verfügbar, deren Angebot ebenfalls Metadaten zum Mikrozensus umfasst[10]: Das Informationssystem GENESIS (Gemeinsames neues statistisches Informationssystem) der Statistischen Ämter des Bundes und der Länder, das „FDZ-Metadatensystem Online" der Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder und das Metadatensystem für arbeitsmarktforschungsrelevante Datensätze des Forschungsinstituts zur Zukunft der Arbeit (IZA). Diese Informationssysteme dienen vor allem dazu, sich einen ersten Überblick über verschiedene Datenquellen zu verschaffen, und beziehen sich dabei auf die on-site Files der Mikrozensen, während das GML sich in seinem Angebot auf die Scientific Use Files (SUF) und deren Nutzer konzentriert. Im Gegensatz zu dem Informationsumfang der drei Systeme wird mit MISSY das Ziel einer möglichst vollständigen Dokumentation verfolgt, die sich nicht auf die unmittelbaren und grundlegenden Metadaten beschränkt, sondern alle für die Analyse der Daten relevanten Informationen umfasst, die nach inhaltlichen Gesichtspunkten verknüpft werden. Die sozial- und wirtschaftswissenschaftliche Ausrichtung der Informationsaufbereitung und -verknüpfung ist eine Besonderheit von MISSY, welche die anderen Informationssysteme in dieser Form nicht leisten können. Daher stellt MISSY eine unverzichtbare Ergänzung zu den anderen Informationssystemen dar.

5 Erfolgte/geplanten Veröffentlichungen des Ergebnisses

MISSY II wurde am 1. März 2010 im Webangebot der GESIS veröffentlicht und ist seitdem frei über das Internet verfügbar.

Publikationen:

Wira-Alam, Andias und Oliver Hopt: Implementing DDI 3.0: a case study of the German Microcensus. 1st Annual European DDI Users Group Meeting: DDI - The Basis of Managing the Data Life Cycle. Bonn, 2009.

[10] Eine Beschreibung dieser Informationssysteme ist bereits im Projektantrag enthalten. Uns sind keine relevanten Weiterentwicklungen bei diesen Stellen bekannt.

Vorträge im Rahmen des Projekts:

Wira-Alam, Andias: DDI Workshop. Dagstuhl 2008.

Hopt, Oliver: 35th IASSIST Conference. Tampere, 2009.

Wira-Alam, Andias: 1st Annual European DDI Users Group Meeting: DDI - The Basis of Managing the Data Life Cycle. Bonn, 2009.

Wira-Alam, Andias: 36th IASSIST Conference. Ithaca (NY), 2009. (wird am 03.06.2010 gehalten)

Poster:

Bohr, Jeanette: MISSY - Information system for the Mikrozensus. Poster auf der European User Conference for EU-LFS and EU-SILC. Mannheim, 2009.

III Anlage: Erfolgskontrollbericht

Beitrag des Ergebnisses zu den förderpolitischen Zielen

Das MISSY-Projekt ist eingeordnet in die förderpolitischen Ziele zur Verbesserung des Datenzugangs und der informationellen Infrastruktur für Daten der amtlichen Statistik und folgt somit den Empfehlungen der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI), insbesondere auch die Nutzbarkeit des Mikrozensus für die Wissenschaft weiterzuentwickeln. Ein Teilaspekt besteht darin, für die Nutzer von Scientific Use Files des Mikrozensus ein datenbankgestütztes Informationssystem zu entwickeln, mit dem Ziel, die bislang zu jedem Scientific Use File zahlreich vorhandenen Metainformationen systematisch zusammenzuführen. MISSY ermöglicht Forschern den schnellen und effizienten Analysezugang zu den Daten und bewirkt damit erhebliche Einsparungen von Zeit und monetären Ressourcen, da die verschiedenen Grundlagenarbeiten nicht von den Einzelforschern selbst erledigt werden müssen.

Wissenschaftlich-technische Ergebnisse des Vorhabens, erreichte Nebenergebnisse und wesentliche Erfahrungen

Die neue Version des Mikrodaten-Informationssystems wurde fertiggestellt und online für die Nutzer verfügbar gemacht. Zum Zeitpunkt des Projektendes sind die kompletten Informationen zu allen Scientific Use Files von 1973 bis 2006 aufbereitet. Neben dem Informationssystem wurde ein Erfassungswerkzeug entwickelt, das auch zukünftig zur Eingabe von Metadaten genutzt werden kann. Das Erfassungswerkzeug unterstützt das Metadaten-Format DDI 3.0 und kann ortsunabhängig genutzt werden.

Fortschreibung des Verwertungsplans

Das in der Projektlaufzeit inhaltlich und informationstechnologisch weiterentwickelte Mikrodaten-Informationssystem MISSY stellt von nun an die Plattform für alle im GML anfallenden Datendokumentationsarbeiten zum Mikrozensus dar und dient als Standard für die Metadatenaufbereitung neu zu dokumentierender Datensätze. Das neu entwickelte Erfassungswerkzeug für variablenbezogene Informationen und die Typo3-Installation zur Pflege von Webseiten und Dateien bieten dabei leicht handhabbare Tools für die einheitlich strukturierte Aufbereitung und Erfassung der relevanten Metadaten. Für eine Nachnutzung des DDI 3.0-Editors ist noch eine Fortführung der Entwicklungsarbeit notwendig, eine Nachnutzung ist jedoch grundsätzlich vorgesehen.

Zudem konnte die Kooperation mit der Mikrozensus Fachabteilung des Statistischen Bundesamtes mit dem Ziel einer effizienteren Datenaufbereitung ausgebaut werden. Als Ergebnis dieses Austauschprozesses werden die in MISSY entwickelten Konventionen zu Variablen- und Valuelabels mittlerweile bereits in der Aufbereitungsphase der Mikrozensus Scientific Use Files im Statistischen Bundesamt umgesetzt und im XML-Format geliefert.

Einhaltung der Ausgaben- und Zeitplanung

Nicht umgesetzt werden konnten die geplanten Arbeiten zur Einbindung der Metadaten des Mikrozensus Regionalfiles, da das File von den Statistischen Ämtern nicht – wie ursprünglich angekündigt –

innerhalb des Projektzeitraums für die Forschung zur Verfügung gestellt wurde[11]. Die für diese Arbeiten eingeplanten Ressourcen wurden dazu genutzt, die Metadaten zu den Mikrozensus Scientific Use Files 2005 und 2006 vollständig in MISSY einzubinden, so dass das Angebot von MISSY bei der Veröffentlichung Metadaten zu allen verfügbaren Mikrozensus Scientific Use Files umfasst.

Die Veröffentlichung des neuen Angebotes hat sich aus technischen Gründen um einige Wochen verzögert, da sich die Erzeugung von sehr langen Listen in DBClear als nicht ausreichend performant erwiesen hatte. Das Problem konnte durch den Einsatz eines Cache-Servers gelöst werden.

Die ursprünglich beantragte Integration von MISSY in Sowiport ist nicht mehr notwendig. Konkret war geplant, die im Projekt erstellten Metadaten aus einem System heraus in zwei Portalen (MISSY und Sowiport) mit jeweils eigenem Layout anzuzeigen und je nach Möglichkeiten unterschiedlich zu vernetzen. Eine Integration von MISSY in Sowiport wird nun aber schon dadurch erreicht, dass Sowiport zukünftig stärker in das Webangebot von GESIS integriert wird.

[11] Das Mikrozensus Regionalfile ist seit Januar 2010 als Scientific Use File verfügbar.

IV Kurzfassung: Berichtsblatt

1. ISBN oder ISSN	2. Berichtsart (Schlussbericht oder Veröffentlichung) Schlussbericht
3. Titel Abschlussbericht zum Anschlussvorhaben Forschungsverbund Datenzentren. Verbesserung des Zugangs der Wissenschaft zu Mikrodaten. Teilprojekt: Pilotprojekt zum Aufbau eines Servicezentrums für Mikrodaten der GESIS. Mikrodaten-Informationssystem MISSY	
4. Autor(en) [Name(n), Vorname(n)] Jeanette Bohr, Oliver Hopt, Andrea Lengerer, Julia Schroedter, Andias Wira-Alam	5. Abschlussdatum des Vorhabens 31.12.2009
	6. Veröffentlichungsdatum
	7. Form der Publikation
8. Durchführende Institution(en) (Name, Adresse) GESIS Postfach 12 21 55 68072 Mannheim GESIS Lennéstraße 30 53113 Bonn	9. Ber. Nr. Durchführende Institution
	10. Förderkennzeichen 01UW0707
	11. Seitenzahl
12. Fördernde Institution (Name, Adresse) Bundesministerium für Bildung und Forschung (BMBF) 53170 Bonn	13. Literaturangaben
	14. Tabellen
	15. Abbildungen
16. Zusätzliche Angaben	
17. Vorgelegt bei (Titel, Ort, Datum)	
18. Kurzfassung Die Weiterfinanzierung des Projekts Mikrodaten-Informationssystem MISSY hatte die volle Implementierung des System zur Aufgabe und umfasste die Einbindung von Metadaten zu allen derzeit verfügbaren Scientific Use Files des Mikrozensus-Grundfiles von 1973 bis 2006. Neben dem inhaltlichen Ausbau des Systems ist als weiterer Aufgabenschwerpunkt die informationstechnologische Neuimplementierung des Systems zu nennen, welche auch die Entwicklung eines Erfassungswerkzeuges für DDI-konforme Metadaten beinhaltete. Aufgrund der in den letzten Jahren stark gestiegenen Anzahl an verfügbaren Scientific Use Files bestand das Ziel darin, die zunehmende Komplexität des Datenmaterials für den Nutzer effektiver handhabbar zu machen und zugleich eine Plattform für die Integration weiterer wichtiger Erhebungen aufzubauen.	
19. Schlagwörter Mikrodaten-Informationssystem, Mikrozensus, Metadaten, Scientific Use File	
20. Verlag	21. Preis