Professur für Demografie
Professorship of Demography

Discussion Papers

No. 17/2013

17

Knut Wenzig

Die Metadatenstruktur des Nationalen Bildungspanels

Otto-Friedrich-Universität Bamberg
University of Bamberg

Die Metadatenstruktur des Nationalen Bildungspanels*

Knut Wenzig

Inhaltsverzeichnis

* Für Anmerkungen und Hinweise zu früheren Versionen danke ich Eva Akins, Tobias Koberg, Gerhard Kraft, Christian Matyas und Rebecca Wirries.

Abbildungsverzeichnis

1 Fähigkeiten und konzeptionelle Idee der NEPS-Metadaten

Die Metadaten des Nationalen Bildungspanels (NEPS) erlauben heute die zweisprachige Dokumentation von Erhebungsinstrumenten (z. B. Fragebögen) und Scientific Use Files (SUFs). Sie sind auch integraler Bestandteil der Datenaufbereitung und neuartiger Rechercheangebote. Dieser Text gibt einen Einblick in den Aufbau und die aktuelle Nutzung der NEPS-Metadateninfrastruktur.

1.1 Leistungen

Der erste erreichte Meilenstein des Metadatenprojekts im Nationalen Bildungspanel war die strukturierte Erfassung der Erhebungsmaterialien, also der Fragebögen nebst Kodierblättern und Programmiervorlagen. Auf dieser Grundlage konnte ein Übersetzungsprozess aufbauen, in dem die Übersetzungsleistungen der externen Agenturen wiederverwendet werden können. Die Erhebungsmaterialien stehen heute den Datennutzer_innen des NEPS u. a. als generierte PDF-Dokumente zur Verfügung.

Weiterhin werden in den NEPS-Metadaten die bereitgestellten Scientific Use Files (SUFs) nahezu komplett abgebildet. Diese Informationen steuern auch die Aufbereitung der SUFs und spielen damit in diesem Bereich eine entscheidende Rolle: Die Variablennamen können für ein internationales Publikum angepasst werden, die Reihenfolge der Variablen in Datensätzen wird durch die Metainformationen gesteuert und in den Metadaten nicht enthaltene Variablen werden nicht ausgeliefert. Auch englische Versionen der Datensätze werden mit Hilfe der Metadaten erzeugt.

Zwischen den Metainformationen für (Erhebungs-)Instrumente und SUFs besteht auf Variablenebene eine Verknüpfung. Mit diesen Informationen können beispielsweise in Stata-Datensätzen die zu Variablen zugehörigen Fragen aus dem Erhebungsinstrument angezeigt werden. Auf dieser Grundlage werden auch zweisprachige Codebooks für die Daten erzeugt, die neben den Fragen zu einzelnen Variablen auch Häufigkeitsauszählungen erhalten und ohne Zugriff auf die Daten einen ersten Einblick in das Datenmaterial ermöglichen.

Mit der frei zugänglichen Webanwendung NEPSplorer steht schließlich eine Benutzerschnittstelle zu den Metadaten zur Verfügung, die das gesamte Datenangebot des Nationalen Bildungspanels per Mausklick erschließt.

Wie die einzelnen Bausteine der NEPS-Metadateninfrastruktur strukturiert sind und ineinandergreifen, wird in diesem Papier ausführlich beschrieben. Sie würden nicht funktionieren, wenn nicht viele Mitarbeiterinnen und Mitarbeiter im NEPS disziplinübergreifend zusammenwirken:

Als erfahrener Softwareentwickler ist Gerhard Kraft für das Anforderungsmanagement an die Software verantwortlich; er steuert für viele Komponenten die Programmierung.
Manfred Dusold herrscht über die SQL-Datenbank und damit über eine Infrastruktur, die nur selten wahrgenommen wird. Die extrem kurzen Reaktionszeiten des NEPSplorers sind wahrscheinlich der sichtbarste Ausdruck seiner Kenntnisse; für diesen Text hat er die Statistiken aus der Datenbank ermittelt.
Christian Matyas ist für die Programmierung des NEPSplorers verantwortlich. Er hat eine Webanwendung entstehen lassen, die eine gezielte Suche innerhalb aller erfassten Metadaten ermöglicht und sich auch im internationalen Vergleich sehen lassen kann.
Als der einzige Sozialwissenschaftler im Metadatenteam musste der Autor dieses Papiers die inhaltlichen Anforderungen so formulieren, dass sie auf der technischen Seite verstanden wurden.

Teile der Programmierung wurden durch das DIPF, einem NEPS-Konsortialpartner ausgeführt: Ralph Köhler zeichnet verantwortlich für die Ausarbeitung von verschiedenen Berichten, die die Daten aus der Datenbank wieder in eine menschenlesbare Form bringen. Dies sind zum Beispiel die Ausgabe von Fragebögen und Programmiervorlagen, die, obwohl in den Metadaten keine Layoutinformationen abgelegt werden, den Originalmaterialien sehr ähnlich sind.
Gerald Mahlmeister ist als externer Programmierer verantwortlich für den Metadateneditor, über den die Informationen in die Datenbank eingegeben werden können.
Eugenia Tsoukrova und Andrea Hesse haben ihren Anteil an vielen Eigenschaften des NEPSplorers.

Der Aufbau der Metadatenstruktur erfordert Kenntnisse, die in den Sozialwissenschaften zunächst nicht verfügbar sind, weil Software entwickelt werden muss. Dieser Entwicklungsprozess selbst verlief nicht immer geradlinig. Einige Entwicklungsziele mussten gestrichen, andere neu integriert werden. Bisweilen prallten auch Sozialwissenschaften und Informationstechnologie aufeinander. Das Ergebnis jedoch kann sich sehen lassen: Die NEPS-Metadateninfrastruktur mit ihrer ausgeprägten vertikalen Integration ist eine Grundlage für die Einhaltung der knappen Zeitpläne bei der Veröffentlichung von Daten für die wissenschaftliche Öffentlichkeit.

Viele Kolleginnen und Kollegen – nicht zuletzt viele studentische Hilfskräfte – sind an der Erfassung der Metadaten beteiligt. Ohne ihre Anstrengungen würden Metadaten und darauf aufbauende Angebote nicht in diesem Umfang zur Verfügung stehen.

1.2 Vertikale Integration: Wiederverwendung als Paradigma

Durch die strukturierte Erfassung der Metadaten können an vielen Stellen Inhalte und auch Strukturen selbst wiederverwendet werden. Damit werden einerseits Ressourcen eingespart, wenn wiederholte Fragen nicht erneut eingegeben und damit auch nicht mehr übersetzt werden müssen. Andererseits wird aber auch die Konsistenz gesichert, weil vor der Neuformulierung von Fragen für eine Erhebung eine leistungsfähige Datenbank für die Recherche bereits erfasster Formulierungen zur Verfügung steht.

Wiederverwendung findet insbesondere an folgenden Stellen statt:

- Ausprägungen von Variablen, sog. Schemata, werden nur einmal erfasst und ggfs. wiederverwendet, auch ihre Übersetzungen.
- Soweit möglich, werden bereits erfasste Fragen in weiteren Fragebögen wiederverwendet.
- Übersetzte Variablen- und Wertelabel aus Programmiervorlagen und Kodierblättern werden, soweit zutreffend, auch in den Datensätzen verwendet.
- Aus der Sicht der Datenbank ist die Struktur der Datensätze und Instrumente (z. B. Fragebögen) weitgehend identisch. Der Übersetzungsprozess für die Label in den Datensätzen konnte unmittelbar übernommen werden.
- Da die Metadaten strukturiert und normalisiert abgelegt sind, können sie ganz unterschiedlich dargestellt und wiederverwendet werden: Als generierte CATI-Programmiervorlage, PAPI-Fragebogen oder als Codebook, in NEPSplorer oder in Datensätzen. So muss eine Korrektur in einem Variablenlabel aus einer Programmiervorlage nur einmal vorgenommen werden und wird dann auch in Datensatz und NEPSplorer erscheinen. Es ist auch ausgeschlossen, dass ein Text in den einzelnen Darstellungsweisen fälschlicherweise voneinander abweicht, weil ein manueller Eingriff überflüssig ist.

Strukturierung und Wiederverwendung stellen auch Ansprüche an alle Beteiligten: Eine explizite Modellierung ist in den meisten Fällen erforderlich, Ausnahmen von der strukturierten Form können i. A. nicht mehr gemacht werden.

2 Kernbereiche der Metadatenstruktur

2.1 Abbildung der Erhebungsinstrumente

Im NEPS werden Erhebungen fast ausschließlich mit Papierfragebögen (PAPI) oder computerunterstützt (CATI/CAPI) durchgeführt. Diese Erhebungsinstrumente, die von Säulen und Etappen erstellt werden, sind Grundlage für die Datenerhebung durch die Erhebungsinstitute. Neben den Frageformulierungen müssen also auch Variablennamen und Variablenlabel vorgegeben werden, damit die zu liefernden Datensätze ausreichend spezifiziert werden. Bei Papierfragebögen sind dies zwei getrennte Dokumente: Einerseits der Fragebogen in der Druckfassung, andererseits das Kodierblatt, in dem Variablennamen, -label, Werte mit Labels und die Zuordnung von Antworten zu Zahlenwerten festgelegt wird. Bei Programmiervorlagen (CATI) sind beide Informationen integriert und im Erhebungsinstitut werden auf dieser Grundlage

die Befragungen programmiert, die dann computerunterstützt durchgeführt werden. Die Ausgangsmaterialien werden hierzu mit gängiger Office-Software in einem mehrstufigen Redaktionsprozess, an dem das ganze Konsortium beteiligt ist, erstellt.[1]

Grundsätzlich unterscheiden sich die beiden Erhebungsmodi also nicht in den Anforderungen an die Datenhaltung. Von außen nach innen ergibt sich folgende Gliederung, die sich auch in der Abbildung 1 wiederfindet:

- Erhebungsinstrument mit Titel und einem einleitenden und abschließenden Text,
- Kapitel oder Module in beliebiger hierarchischer Gliederung, mit jeweils einer Überschrift und einem einleitenden und abschließenden Text,
- Fragen,
- Fragenummer und Filterinformationen – als Eigenschaften der Frage, wenn sie in diesem Instrument verwendet wird.

Hierbei besteht eine Frage aus

- einer Fragetext (ggfs. mehrere Paare aus Bedingung und Fragetext),
- einer Intervieweranweisung für verschiedene Befragungsmodi sowie
- einer oder mehreren Variablen, zu der ein Variablenname, ein Variablenlabel, eine Teilfrage, eine Antwortvorgabe (s. u.), ggfs. (meist bei CATI-Interviews) fehlende Werte (z. B. keine Angabe, weiß nicht) und Schemaerweiterungen (das sind weitere Paare von Werten und Wertelabels) gehören.

Als Antwortvorgabe kann neben einem Freitext oder einer numerischen Angabe auch ein Antwortschema gewählt werden. Es besteht aus einer geordneten Liste von Paaren aus einem numerischen Wert und einem Text, dem Wertelabel. Die Datenbank ist hierbei so strukturiert, dass Antwortschemata in verschiedenen Variablen beliebig wiederverwendet werden können (und sollen). Fragen können in verschiedenen Instrumenten wiederverwendet werden.

Die Abbildung 2 zeigt eine Beispielfrage, die einmal in einer CATI-Programmiervorlage und einmal in einem PAPI-Fragebogen verwendet wird. Folgendes fällt auf:

- Eine Frage kann in unterschiedlichen Erhebungsinstrumenten und -modi wiederverwendet werden.
- Nach der Erfassung ist es möglich, Darstellungsformen zu wählen, die unterschiedliche Erhebungsmodi zeigen.

1 Anfangs sollte auch der Entwicklungsprozess der Erhebungsinstrumente unterstützt werden. Damit hätte, was konzeptionell sicherlich bestechend ist, die Metainformation direkt an der Quelle erfasst werden können. Wegen der Komplexität der sich hierbei stellenden Anforderungen wurde dieses Ziel jedoch fallengelassen und man verständigte sich darauf, vorerst die Erhebungsinstrumente nur so zu dokumentieren, wie sie im Feld eingesetzt wurden.

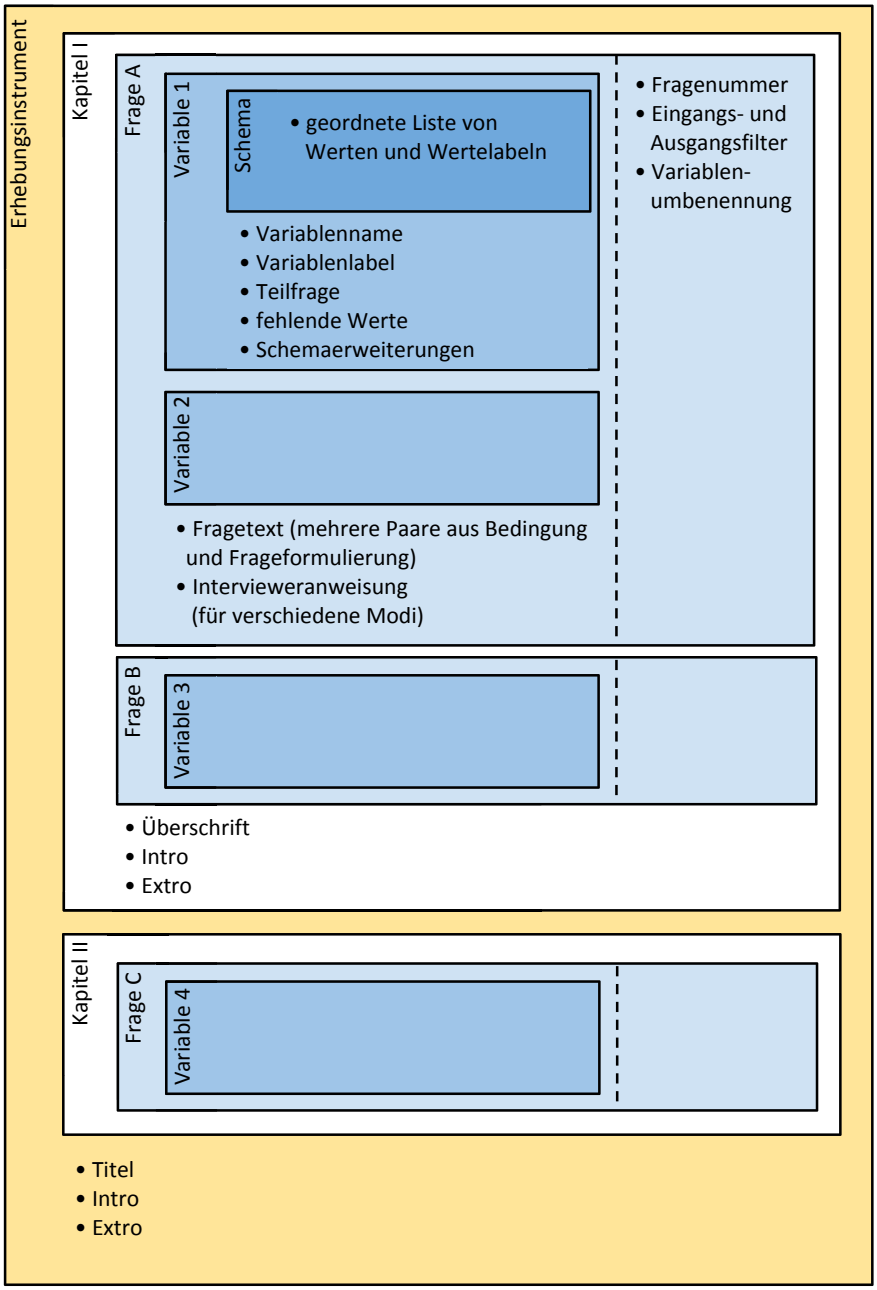

Abbildung 1: Schematischer Überblick über die in den NEPS-Metadaten gespeicherten Informationen.

22203	--va: **etzeit** --fn: **22203** --vb: Zeitarbeit --fr: Waren Sie da als Zeitarbeiter/in oder Leiharbeiter/in beschäftigt? --in: <<auch in Personalserviceagentur>> --we: 1: ja 2: nein **BUTTONS: Angabe verweigert (-97), Weiß nicht (-98)** --af: **IF 22203 = 1 GOTO 23300** --end--

12	**Waren Sie da als Zeitarbeiter/in oder Leiharbeiter/in beschäftigt?**
ja	☐
nein	☐
Angabe verweigert	☐
Weiß nicht	☐
"ja": Bitte weiter mit Frage 15	

Abbildung 2: Beispiel für eine Frage, die einmal in einer CATI-Programmiervorlage und einmal in einem PAPI-Instrument verwendet wird.

- Elemente wie Fragenummern oder Filter sind Eigenschaften, die die Frage bekommt, wenn sie in einem bestimmten Instrument verwendet wird. Sie können für die gleiche Frage von Instrument zu Instrument unterschiedlich sein.
- Die Intervieweranweisung fehlt beim PAPI; sie ist modus-spezifisch abgelegt.
- Bei der PAPI-Darstellung fehlen Variablenname und -label sowie die Zahlenwerte für die Ausprägungen. In einer anderen Darstellungsform könnten diese Informationen eingeblendet werden, um etwa die Arbeit mit den Daten zu unterstützen.

Das Beispiel in Abbildung 2 hat gezeigt, dass bei der Verwendung einer Frage in einem Instrument zusätzliche Eigenschaften (Fragenummer und Filter) gespeichert werden können. Auch der Variablenname kann fragebogenspezifisch überschrieben werden. Wenn sich zwei Fragen nur durch die Variablennamen unterscheiden, brauchen sie nur einmal erfasst werden.

Nachdem die Fragen eingegeben und in einem Fragebogen verwendet wurden, kann dieser Fragebogen ausgegeben werden. Damit ist die grundlegendste Anforderung erfüllt: Die bei der Eingabe aufgespaltenen und strukturierten Informationen können wieder in einer Form ausgegeben werden, die dem Rohmaterial sehr nahe kommt.[2]

Die Ausgaben werden durch sogenannte Reports erzeugt, die mit Hilfe der SQL Server Reporting Services von Microsoft zur Verfügung gestellt werden. Mit dieser Funktionalität können sowohl die CATI- als auch die PAPI-Instrumente für Dokumentationszwecke hinreichend gut nachgebildet werden. Während die reine Nachbildung der Originalvorlagen die Kontrolle der Eingaben erleichtert, sind grundsätzlich auch generischere Ausgaben denkbar: Die Darstellung im NEPSplorer unterscheidet nicht nach Erhebungsmodus.

In Programmiervorlagen werden an vielen Stellen Bedingungen verwendet: bei der Filterung des Frageablaufs (vgl. Abbildung 2), der Formulierung der Fragen selbst und auch der Formulierung von möglichen Antwortvorgaben. Was die bedingten Frageformulierungen betrifft, wird das Metadatenschema gerade ergänzt werden, um die Wiederverwendbarkeit von Fragen zu verbessern. Ein grundsätzliches Problem stellen jedoch die Bedingungen selbst dar: Ihr syntaktischer Aufbau ist vielfach nicht so konsistent, dass sie unmittelbar durch eine Software interpretierbar sind, obwohl dies auch bei der Entwicklung und Testung der Instrumente vorteilhaft wäre.

Derzeit (nach Release von SC5 3.0.0) sind 43 Erhebungsinstrumente in der Metadatenbank erfasst. Sie bestehen aus insgesamt 5081 Fragen, davon sind 3891 Fragen einmalig. In 11228 Variablen, die in diesen Fragen definiert werden, werden 840 verschiedene Schemata verwendet.

2 Im Idealfall, in dem sich Rohmaterial und generierte Ansichten gleichen, würde man auf die Archivierung des Rohmaterials an anderer Stelle verzichten und die Datenbank ist der „single point of information“, an dem sämtliche Informationen über das Rohmaterial abgefragt werden können. Dies wird jedoch in der Regel nicht der Fall sein. In den NEPS-Metadaten sind beispielsweise keine Textauszeichnungen oder andere Layout-Informationen abgespeichert.

2.2 Organisation des Übersetzungsprozesses

Mit Blick auf die internationale Scientific Communitiy werden alle Textinformationen in den Metadaten übersetzt. Um den Übersetzungsaufwand zu minimieren und eine gewisse Konsistenz der Übersetzung sicherzustellen, werden einzelne Textbestandteile nur einmal übersetzt und in der Folge wird die Übersetzung wiederverwendet. Die Wiederverwendung der Übersetzungen folgt der Wiederverwendungslogik der Metadaten selbst: Jedes Antwortschema wird nur einmal übersetzt.[3] Dies gilt auch für alle Texte in Fragen. Alle Texte in den Erhebungsinstrumenten selbst (z. B. Überschriften und Einleitungstexte von Abschnitten) müssen jeweils neu übersetzt werden, auch wenn sie in einem anderen Instrument bereits verwendet wurden.

Als Schnittstelle für die Übersetzungsagenturen wird XLIFF verwendet, ein XML-Format, das neben den Textpaaren in Ausgangs- und Zielsprache auch in bestimmtem Umfang Informationen über den Workflow abbilden kann. Innerhalb der XLIFF-Dateien werden die Textpartikel weitgehend in der Reihenfolge angeordnet, wie sie auch im Erhebungsinstrument vorkommen. Damit kann trotz der Abspeicherung der Texte in unterschiedlichen Tabellen der Datenbank der Zusammenhang für den Übersetzungsprozess in gewissem Umfang erhalten werden.

Abbildung 3 zeigt für ein kurzes Instrument mit nur einer Frage eine komplette XLIFF-Datei inklusive der bereits übersetzten Partikel. Das Schema („ja“/„nein“) wurde bereits übersetzt. Auch für die Texte der Frage selbst liegt bereits eine Übersetzung vor, etwa für das Variablenlabel „Zeitarbeit“. Nicht übersetzt wurde bisher nur der Kapiteltitel („Kapitel A“) und der Filter („Falls ja, bitte weiter mit Frage 15“). Das sind auch die beiden Textpartikel, die exportiert würden, wenn man nur die noch zu übersetzenden Bestandteile anfordert.

Sämtliche bisher übersetzten Texte können instrumentübergreifend als XLIFF-Datei exportiert werden. Sie können dann, geeignete Software vorausgesetzt, den Übersetzungsvorgang unterstützen und die Konsistenz der Übersetzung erhöhen. Beispielsweise kann die freie Software Virtaal[4] die übersetzten Texte in ihr „Translation Memory“ integrieren und hieraus Vorschläge für neu zu übersetzende Texte generieren. In Abbildung 4 ist zu sehen, wie Virtaal einen Übersetzungsvorschlag anzeigt.

Die Übersetzungsdatenbank enthält bisher über 24.550 übersetzte Textpartikel mit insgesamt über 174.262 Worten.

2.3 Abbildung der scientific use files

Nicht zuletzt, weil die Datensätze selbst auch zweisprachig angeboten werden sollten, wurde, um den Übersetzungsprozess nutzen zu können, eine Dokumentation der Datensätze in

3 Alternativ wäre es auch möglich gewesen, jedes Textfragment unabhängig von seiner Verwendung immer nur einmal zu übersetzen. Es schien allerdings wünschenswert, kontextabhängig übersetzen zu können, was nun im Rahmen der Wiederverwendungslogik von Schemata und Fragen gewährleistet wird.

4 `http://sourceforge.net/projects/translate/`

```
<?xml version="1.0" encoding="utf-8"?>
<xliff version="1.0">
  <file datatype="plaintext" original="MultiLingualeContent.fla" source-language="de" target-language="
  en">
    <body>
      <trans-unit id="Chapters|Name|ac831872-1a72-412a-b8c2-ac747f39c891">
        <source>Kapitel A</source>
        <target>Kapitel A</target>
        <note>2013-02-04 17:44:22</note>
      </trans-unit>
      <trans-unit id="Questions|Text|6f379ef5-3e92-46e1-bfe9-c20bc7c781af">
        <source>Waren Sie da als Zeitarbeiter/in oder Leiharbeiter/in beschäftigt?</source>
        <target>Did you work as a temporary or contract worker then?</target>
        <note>2011-04-15 08:12:24</note>
      </trans-unit>
      <trans-unit id="Variables|Label|a9e6b874-3184-4e1d-9a71-1e94b42af990">
        <source>Zeitarbeit</source>
        <target>Temporary employment</target>
        <note>2011-04-15 08:12:24</note>
      </trans-unit>
      <trans-unit id="SchemeOptions|Label|6d097ff6-9d91-433f-b7b9-85c589825224">
        <source>ja</source>
        <target>yes</target>
        <note>2010-12-02 12:12:02</note>
      </trans-unit>
      <trans-unit id="SchemeOptions|Label|7dc3da4a-19ed-4038-a68e-ffe21bc8507a">
        <source>nein</source>
        <target>no</target>
        <note>2010-12-02 12:12:02</note>
      </trans-unit>
      <trans-unit id="ConditionalWorkflows|Condition|27d33be3-41e5-4f76-a9ed-e0718f056313,2">
        <source>"ja": Bitte weiter mit Frage 15</source>
        <target>"ja": Bitte weiter mit Frage 15</target>
        <note>2013-02-04 17:47:11</note>
      </trans-unit>
    </body>
    <note>2013-02-04 17:47:11</note>
  </file>
</xliff>
```

Abbildung 3: Eine komplette XLIFF-Datei für das PAPI-Instrument mit einer Frage aus Abbildung 2

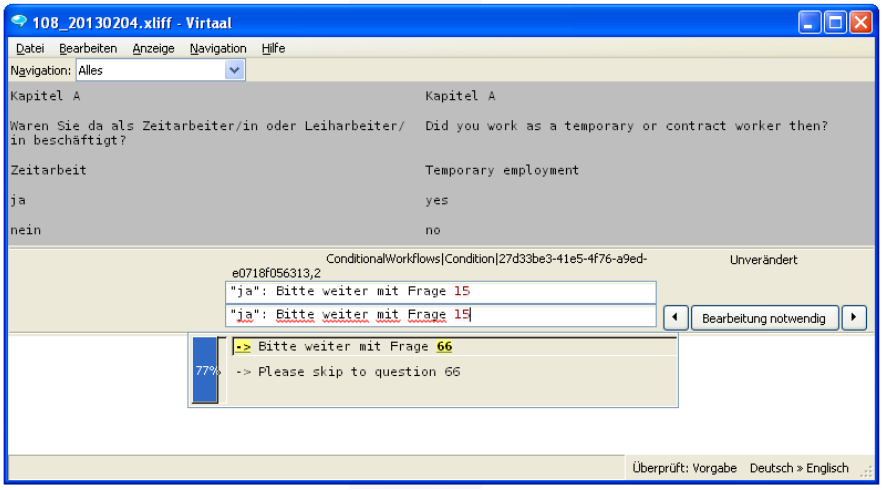

Abbildung 4: Für die erstmalige Übersetzung von Texten können Vorschläge auf Grundlage der bisherigen Übersetzungen gemacht werden.

den Metadaten angestrebt. Der erste konzeptionelle Versuch, die abgespeicherten Instrument-Informationen in zusätzlichen Feldern innerhalb der Datenstruktur des zugehörigen Instruments abzuspeichern, erwies sich hierbei als nicht zielführend. Obwohl die Datensätze mit Hilfe der Erhebungsinstrumente erzeugt wurden, sind die inhaltlichen Unterschiede größer als zunächst vielleicht vermutet.[5] Für eine Dokumentation der SUFs genügt es also nicht, die Erhebungsinstrumente zu erfassen, vielmehr müssen auch die SUFs selbst abgebildet werden.

Schnell stellte sich aber heraus, dass die abstrakte Struktur von einem Erhebungsinstrument und einem aus mehreren Datensätzen bestehenden SUF nahezu identisch ist. Lediglich die Ebene der Frage ist überflüssig, was bedeutet, dass im SUF jede Frage nur mit einer Variablen gefüllt wird.

Die Übersicht in Abbildung 5 zeigt stark vereinfacht, welche Objekte die Datenstruktur enthält und wie sie für Erhebungsinstrumente und SUFs genutzt werden. Ein ausführliches Entity-Relationship-Diagramm findet sich in Abbildung 10 im Anhang.

Abbildung 6 zeigt die Informationsstruktur für einen SUF (rechts) und wie stark sie der eines Erhebungsinstruments (links) ähnelt. Wichtigster Unterschied ist der Umstand, dass Fragen in einem Datensatz nicht vorkommen und damit das Objekt in der Datenbank, in dem Fragen gespeichert werden, zwar verwendet wird, aber immer nur eine Variable enthält. In der Abbildung befinden sich die Variablen 1 und 2 in einer Frage im Instrument, im SUF sind sie aber als Variablen 10 und 11 in getrennten Items (so heißt das Objekt der Frage innerhalb der Datenbank, vgl. Abb. 5) enthalten

5 Es werden nicht nur neue Variablen erzeugt (z. B. im Rahmen einer Skalenbildung), es werden auch Variablen von der Veröffentlichung ausgeschlossen (z. B. Zeitstempel). Auch die Zusammenfassung der Variablen in Datensätzen muss nicht dem Vorkommen in Erhebungsinstrumenten entsprechen.

Objekt in Metadatenstruktur Instrument (Instruments)	Objekt, das dokumentiert wird	
	ein Erhebungsinstrument	gesamter SUF
Kapitel (Chapters)	Kapitel	Dateien
Item (Items)	Fragen	–
Variable (Variables)	ggfs. mehrere Variablen	eine Variable
Schema (Schemes)	Antwortschema	Value labels

Abbildung 5: Die Nutzung von Objekten in der Metadatenstruktur (in Klammern die Namen aus dem ERD, vgl. Abb. 10) zur Dokumentation von Erhebungsinstrumenten und SUFs

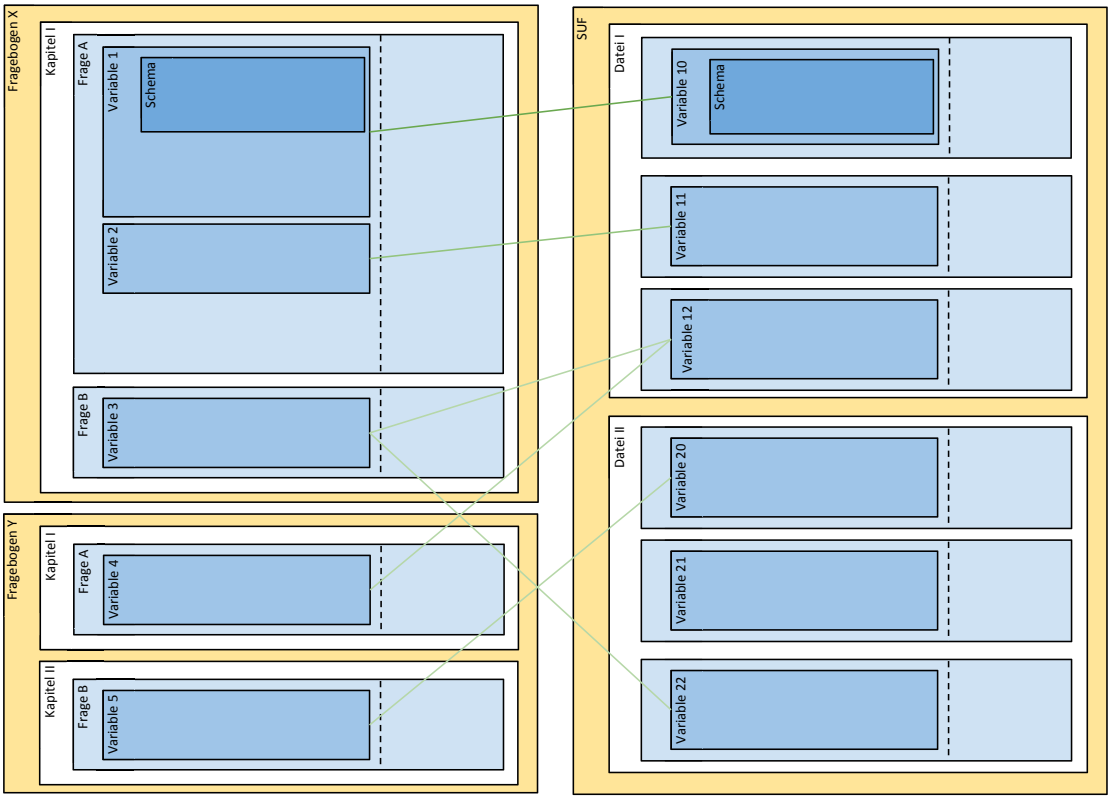

Abbildung 6: Schematische Darstellung der Informationsstruktur von Erhebungsinstrumenten (links) und Scientific Use Files (rechts) sowie deren Verknüpfung.

Auch der bereits angesprochene Zusammenhang zwischen einem SUF und einem oder mehreren Erhebungsinstrumenten, kann und sollte abgebildet werden. Nicht zuletzt wurden mit den Erhebungsinstrumenten die Daten erzeugt, die in den SUF geflossen sind. Dieser Zusammenhang wird in den NEPS-Metadaten über Beziehungen zwischen den Variablen im SUF und im Erhebungsinstrument abgebildet.[6]

In Abbildung 6 wird diese Beziehung durch die grünen Linien dargestellt. Wenn eine solche Beziehung existiert, kann die Frage aus dem Erhebungsinstrument zur Variable im Datensatz zugeordnet werden. Genutzt wird diese Verknüpfung beispielsweise in den Codebüchern.

Dank dieser Referenz erübrigt es sich auch in vielen Fällen, ein neues Variablenlabel zu definieren: Wird im SUF kein Variablenlabel definiert, wird das Variablenlabel der entsprechenden Variable aus dem Erhebungsinstrument übernommen. Ähnlich wird mit dem Antwortschema verfahren. Dies begrenzt die Redundanz in den Daten und ermöglicht auch hier die Wiederverwendung von Übersetzungsleistungen.

3 Nutzung, Erweiterung und Pflege der Metadaten

3.1 Metadatengetriebene Aufbereitung der SUFs

Die in den Erhebungsinstrumenten definierten Variablennamen eignen sich aus vielerlei Gründen nicht zur Veröffentlichung. Häufig wurden z.B. (deutsch) sprechende Variablennamen vergeben, was für einen Datensatz mit internationalem Publikum nicht angemessen ist. Nachdem festgelegt wurde, weitgehend systemfreie (nicht sprechende[7]) Variablennamen (vgl. Abbildung 7) zu vergeben, sollte die Umbenennungsinformation in den Metadaten abgebildet werden und auch durch die Metadaten selbst gesteuert werden.

Diese Information befindet sich im SUF-Objekt der Metadatenbank, das etwa wie folgt erstellt wird:

Zunächst wird hierbei von den jeweils zuständigen Fragebogenentwicklern instrumentweise für jede Variable im Instrument ein neuer Variablenname aus ihrem Namensblock[8] vergeben. Die Information wird als sogenannter SUF-Variablenname im Erhebungsinstrument als Eigenschaft der Frage (die eine oder mehrere Variablen enthält) abgespeichert. Diese Information

6 Hierbei kann eine Variable im SUF aus mehreren Variablen in Erhebungsinstrumenten gebildet werden (z. B. im Panelfall bei Datenhaltung long), wobei eine der Beziehungen als primäre Beziehung qualifiziert werden kann. Die so referenzierte Frage kann dann beispielsweise bevorzugt angezeigt werden.

7 Lediglich die Art der Zielperson und die (zunächst) zuständige Arbeitseinheit ist aus den Variablennamen abzulesen.

8 Die letzten 5 Stellen des 7-stelligen Variablennamens, wobei die letzte Stelle alphanumerisch und alle anderen numerisch sein müssen.

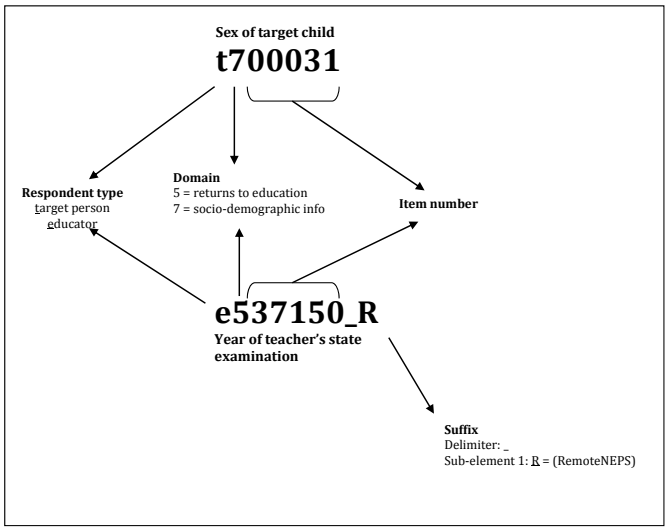

Abbildung 7: Der Aufbau der Variablennamen in den Scientific Use Files des NEPS. (Leopold, Raab und Skopek 2011, S. 8)

wird verwendet, um die SUF-Variablen in den generierten Erhebungsinstrumenten anzuzeigen. Die Aufbereitung des SUF erfolgt, bis die Struktur der Dateien feststeht, noch mit den alten Variablennamen. Dann werden für jede Datei die (alten) Variablennamen ausgegeben. Für diese wird versucht, aus dem Instrument einen neuen Namen zu ermitteln und das SUF-Objekt in den Metadaten wird mit den einzelnen Dateien, neuen Namen und alten Namen initialisiert. Auf dieser Grundlage, also mit Informationen, die ausschließlich im SUF-Objekt verfügbar sind, wird dann die Umbenennung durchgeführt.[9]

Neben den Variablennamen wird auch die Reihenfolge der Variablen durch die Metadaten gesteuert. Weiterhin werden Variablen, die nicht in den Metadaten enthalten sind, im Rahmen der Aufbereitung nicht in die endgültigen Datensätze ausgeliefert.

Nachdem die Struktur der Datensätze vollständig in den Metadaten abgebildet wird, können die (Variablen- und Werte-)Labels in den Datensätzen im Rahmen des Aufbereitungsprozesses durch die entsprechenden Informationen in den Metadaten vollständig ersetzt werden. Dies scheint auf den ersten Blick riskant, ist – weil auch englische Datensätze produziert werden müssen – gleichzeitig alternativlos. Wenn die Labels der deutschen Datensätze mit den Metadaten überschrieben werden, finden mehrstufige computerunterstützte Prüfschleifen statt, um sicherzustellen, dass Texte nicht sinnentstellend überschrieben werden. Diese Prüfschleifen sind aber mit zwei zu vergleichenden deutschen Textpartikeln einfacher zu bewältigen, als mit einem deutsch/englischen Textpaar.

9 Das ist gleichzeitig auch eine Option für eine künftige Weiterentwicklung der Datenstruktur: Umbenennungsinformationen werden teilweise redundant im Bereich des Instruments und des SUFs abgelegt. Es ist denkbar, diese Redundanz aufzulösen und durch eine in den Metadaten gespeicherte Ableitungsvorschrift des Variablennamens zu ersetzen.

Derzeit enthält die NEPS-Metadatenbank Informationen 9 SUFs mit 142 Dateien und 15566 Variablen. Die Versionierung der Datensätze[10] wird in den Metadaten abgebildet, indem die Informationen der Vorversion in ein neues SUF-Objekt kopiert werden.

3.2 Abbildung des Erhebungsdesigns

Die Erhebungsinstrumente können nicht nur als solche dokumentiert werden, vielmehr kann auch ihr Einsatz im Multikohortensequenzdesign (Blossfeld, Maurice und Schneider 2011, S. 14), dem übergreifenden Erhebungsdesign des NEPS, das aus 6 Startkohorten und 2 Zusatzstudien besteht, abgebildet werden. Das entsprechende Datenmodell ist im blauen Bereich der Abbildung 10 zu finden und beruht auf folgenden Überlegungen:

- Es gibt einzelne Stichproben von Zielpersonen, d. h. Neugeborene, Kinder in Kindergärten und Schulen, Studierende oder Erwachsene. (vgl. Tabelle „Samples“ im ERD im Anhang)
- Diese Stichproben werden zu bestimmten Zeitpunkten eingesetzt. (vgl. Tabelle „SampleWaves“)
- An diesen Einsatzzeitpunkten wird entschieden, ob die Zielpersonen selbst und/oder Personen im Kontext (Eltern, Lehrer, Schulleiter_innen) befragt werden. Dies sind die eigentlichen Studien, für die die Erhebungsinstrumente entwickelt werden. (vgl. Tabelle „Surveys“[11])

Zu diesen so definierten Befragungen können die Erhebungsinstrumente zugeordnet werden. (vgl. Tabelle „InstrumentSurveys“)

Weiterhin können die Stichproben selbst in ein hierarchisch strukturierbares System von Gruppen geordnet werden. Auf der obersten Ebene stehen hierbei die 6 Kohorten und 2 Zusatzstudien, auf Ebenen dazwischen beispielsweise die Unterscheidung zwischen Förderschule und Regelschule. (vgl. Tabelle „SampleGroups“ in Abbildung 10) Mittels der Informationen aus dem Erhebungsdesign werden die Erhebungsinstrumente in der Einstiegsmatrix (vgl. Abbildung 8) des NEPSplorers plaziert.

3.3 Metadatenprodukte

Wie bereits geschildert, sind die SUFs als solche bereits Produkte, die ohne die Metadaten nicht mehr denkbar sind. Die Stata-Versionen enthalten jedoch eine ganz besondere Information: Weil die einzelnen Variablen im Datensatz innerhalb der Metadaten direkt mit den Fragetexten in den Erhebungsinstrumenten verknüpft sind, stehen die Texte bei der Aufbereitung der Datensätze auch direkt zur Verfügung und können über die characteristics-Funktionalität von

10 „Um künftige Aktualisierungen der einzelnen SUFs verwalten zu können, verwendet das NEPS-Datenzentrum ein 3-stelliges Versionierungssystem. Die erste Stelle entspricht hierbei in der Regel der Welle (main release), die anderen beiden Stellen werden für major und minor updates verwendet.“ (Wenzig 2012, S. 3)

11 Die hier zugeordnete Tabelle „Studies“ stellt eine Verbindung her zu einem NEPS-internen System von Studiennummern.

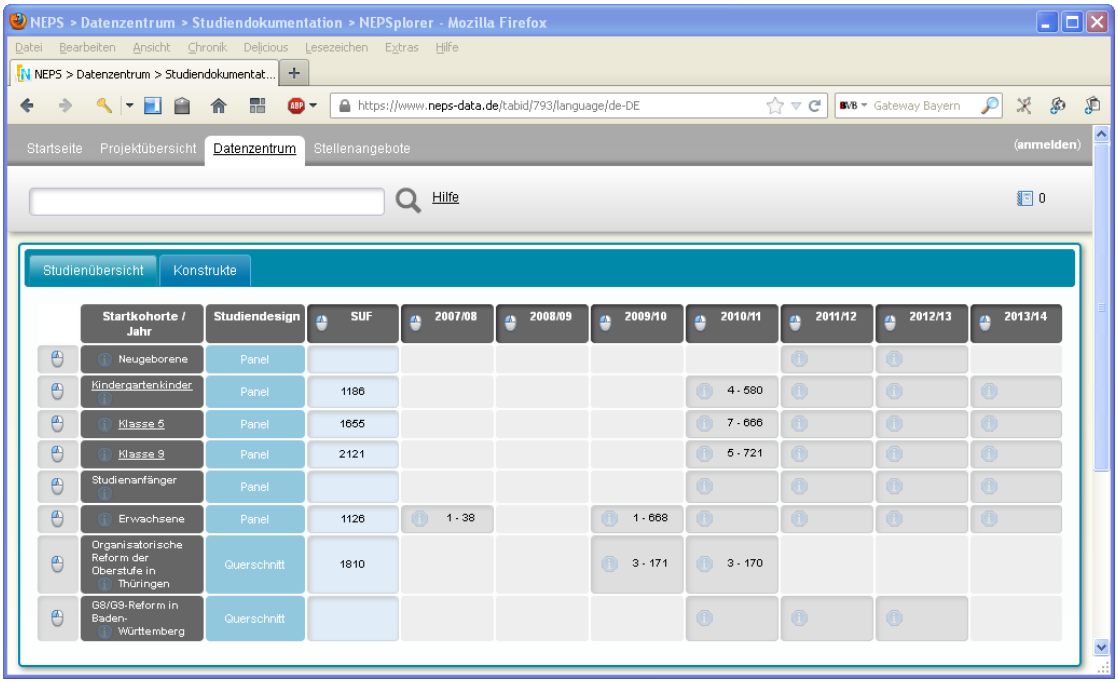

Abbildung 8: Der NEPSplorer startet mit einer vereinfachten Übersicht über das NEPS-Erhebungsdesign.

Stata zu den Variablen im Datensatz abgespeichert werden. Von Daniel Bela, der für die Intergration der Metadaten bei der Datenaufbereitung verantwortlich ist, stammt infoquery, ein Befehl der direkt in Stata die Ausgabe des zur Variablen gehörenden Fragetextes ermöglicht. Ihm ist es auch zu verdanken, dass die Stata-Datensätze zweisprachig ausgeliefert werden. Mit dem Befehl label language kann die Sprachversion umgeschaltet werden.[12]

Das innovativste Metadatenprodukt im NEPS ist sicherlich der NEPSplorer. Er verortet die Ergebnisse einer leistungsfähigen Volltext- und Schlagwortsuche unmittelbar im NEPS-Studiendesign und ermöglicht einen extrem schnellen Zugriff auf Erhebungsinstrumente, Fragen, Datensätze und Variablen aller bereits veröffentlichten Scientific Use Files. In gewissem Sinne eine Druckfassung des NEPSplorers stellen die Codebücher dar, die skriptgesteuert aus den fertigen Datensätzen und den zugehörigen Metadaten zweisprachig erstellt werden. Sie enthalten für jeden SUF, sortiert nach Datensätzen, für jede Variable Häufigkeitsauszählungen mit Variablen- und Wertelabels, sowie der zugehörigen Frage.

Auch die Ansichten der Erhebungsinstrumente, die sich am besten zur Arbeit mit den Daten eignen, werden mit Metadaten erzeugt: Die produzierten Programmiervorlagen sind sehr nah am Original. Größere Unterschiede gibt es, weil keine Layout-Informationen vorliegen, bei den PAPI-Fragebögen. Dafür können in einer den Fragebögen zumindest ähnlichen Ansicht Va-

12 SPSS bietet keine vergleichbare Funktionalität, hier wird jeweils eine deutsche und eine englische Version ausgeliefert.

riablennamen, -label sowie numerische Werte, die den Antworten zugeordnet sind, angezeigt werden. In den Ansichten für beide Instrumentarten können dabei entweder die ursprünglichen oder die im SUF verfügbaren Variablennamen angezeigt werden.

3.4 Anreicherung der Metadaten

Auf Itemebene, also auf Ebene der Fragen in Erhebungsinstrumenten oder der Variablen im Datensatz sollen künftig weitere Informationen erfasst werden, die das Finden von Fragen und Variablen erleichtern, die wissenschaftliche Qualität der Dokumentation erhöhen oder interne Abläufe gestalten helfen.

So sollen künftig möglichst Fragen verschlagwortet werden. Hierzu wurde eine Klassifikation erarbeitet, der NEPS-Konstruktbaum[13]. Die bereits genutzten Schlagworte sind bereits im NEPSplorer (Registerkarte „Konstrukte“) sichtbar und können für die Suche genutzt werden. Derzeit wird geprüft, wie weit eine Verknüpfung zum „Thesaurus Soziologie“ möglich ist und ob hierdurch die Suche verbessert werden kann.

Weiterhin sollen für einzelne Fragen Quellenangaben erfasst werden. Auf diese Weise werden nicht nur Querbeziehungen zu anderen Studien zuverlässig hergestellt, auch der Anspruch „Give credit where credit is due.“ kann eingelöst werden.[14]

Für die interne Nutzung, auch zur Steuerung von Arbeitsabläufen sollen weiterhin zu einzelnen Fragen die verantwortliche Arbeitseinheit und eine Ansprechperson erfasst werden. Dies ermöglicht die Erstellung von entsprechenden Listen, was die Itementwickler_innen entlasten könnte, weil sie keine eigene Dokumentation mehr vorhalten müssen.

3.5 Pflege der Metadaten

Zur Erfassung der Metadaten wurde mit der Programmierung einer speziellen Software begonnen, dem Metadateneditor. Er greift auf eine zwischenzeitlich recht komplex gewordene SQL-Datenbank (vgl. Abbildung 10) zu und ermöglicht die Pflege nahezu aller Informationen aus den Erhebungsinstrumenten. Der grobe Aufbau des user interface folgt der Wiederverwendungslogik der Metainformationen. Es können Schemata erfasst werden, die in Variablen von Items verwendet werden können. Die Items werden dann in (Erhebungs-) Instrumenten angeordnet. Unter „SUF“ wird eine Ansicht auf die Instrumente der Datenbank vermittelt, die die Scientific Use Files abbilden. Hier ist es u. a. möglich, die Verknüpfung einzelner Variablen im SUF zu Variablen in Erhebungsinstrumenten zu erfassen.

13 Eine dynamisch erzeugte Version des vollständigen Baums findet sich unter: `https://www.neps-data.de/de-de/datenzentrum/studiendokumentation/konstruktbaum.aspx`.

14 Nach derzeitigen Planungen wird hierbei auf die Bibliographie-Plattform „BibSonomy“ (http://www.bibsonomy.org) zurückgegriffen. Sie ermöglicht kollaborative Erfassung der bibliographischen Informationen und unterstützt durch die Nähe zu BibTEX die automatische Erstellung von Dokumenten. Weiterhin können Forscherinnen und Forscher auf diese Datenbank selbst zugreifen, womit das NEPS auch an dieser Stelle einen Beitrag zur Verbesserung der wissenschaftlichen Infrastruktur leistet.

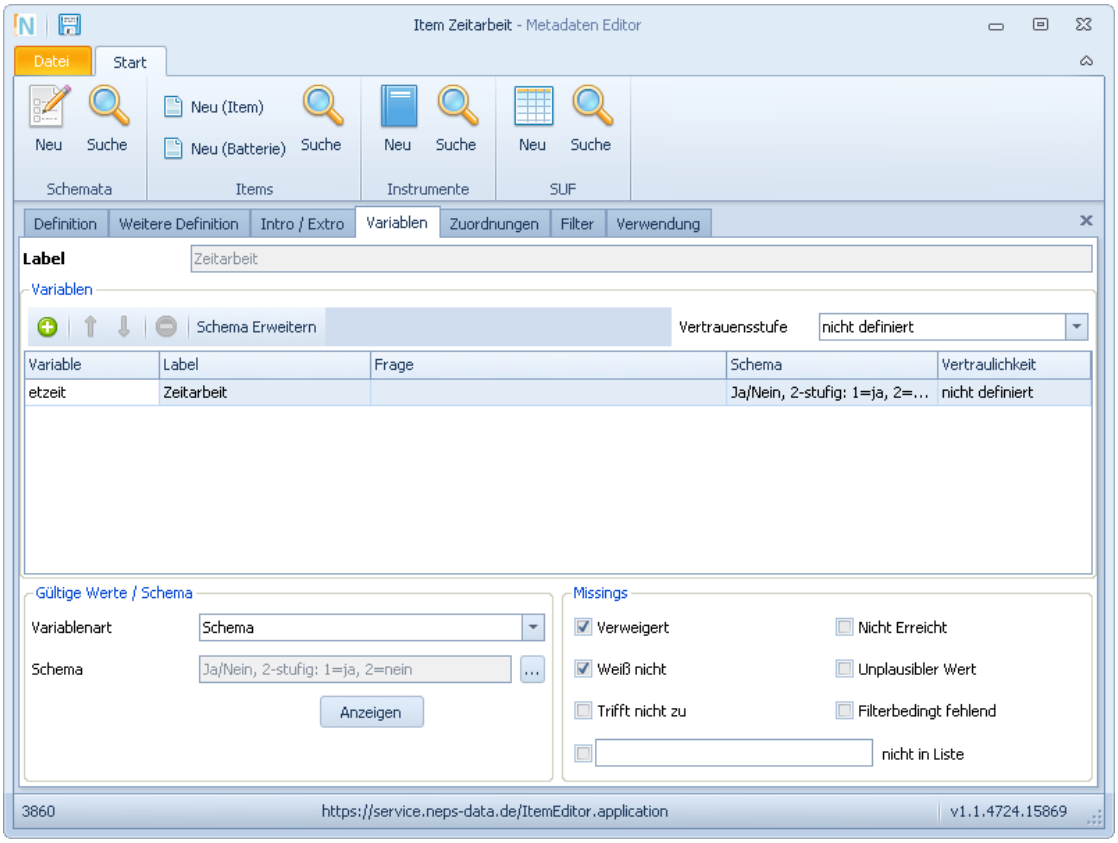

Abbildung 9: Der Metadateneditor ermöglicht den Zugriff auf kleineste Informationseinheiten in den Metadaten.

Große Mengen von Metadaten, die bereits in einem automatisch verarbeitbaren Tabellenformat abgelegt sind, werden über halb standardisierte Schnittstellen in den Datenbestand importiert. Dies wird immer dann genutzt, wenn verfügbare Informationen schon so vorliegen, dass der Importaufwand für den Datenbankspezialisten gegenüber die manuellen Erfassung durch Hilfkräfte gering erscheint. Ein Instrument wurde auch mittels einer XML-Datei erzeugt, die mit geringem manuellen Eingriff aus einer Wordprogrammiervorlage erzeugt wurde.

4 Ausblick

Derzeit bildet die NEPS-Metadatenbank die Informationen so ab, dass eine Dokumentation der veröffentlichten Datensätze möglich wird. Die Datenstruktur hat sich weitgehend als sehr nützlich erwiesen. Weiterentwicklungen bieten sich in folgenden Bereichen an:

Schnittstelle zu DDI DDI, insbesondere in der Variante Lifecycle, ist ein XML-basierter Standard mit dem Anspruch, Daten zu dokumentieren und Verarbeitungsprozesse zu steuern. Um von Vorarbeiten profitieren zu können, die in diesem Kontext weltweit geleitet werden, wäre eine zu entwickelnde Schnittstelle wünschenswert.

Verarbeitungsprozesse dokumentieren und steuern Die Generierung etwa von neuen Variablen findet derzeit unabhängig von den Metadaten im Aufbereitungsprozess statt. Hier bietet sich zunächst an, die Skripte der Aufbereitung variablenweise mit den Metadaten zu verbinden, um die Dokumentation zu verbessern und den Dokumentationsaufwand gleichzeitig gering zu halten. In einem nächsten Schritt könnte die Aufbereitung durch die Metadaten selbst gesteuert werden, etwa auch im Bereich der Anonymisierung.

Wiederverwendung der Fragen stärken Die Wiederverwendung von Fragen in anderen Erhebungskontexten sollte ermutigt werden, um wissenschaftliche Anschlussfähigkeit zu sichern. Eine öffentlich zugängliche Fragedatenbank kann dazu beitragen, dass andere Forscherinnen und Forscher Fragen aus dem NEPS für eigene Erhebungen nutzen. Ausgangspunkt für ein solches Projekt wäre etwa eine Möglichkeit zum Download von einzelnen Fragen in unterschiedlichen Formaten (z. B. auch DDI).

5 Literatur

Blossfeld, Hans-Peter, Jutta von Maurice und Thorsten Schneider (2011). „The National Educational Panel Study: need, main features, and research potential". In: *Zeitschrift für Erziehungswissenschaft* Special Issue 14: *Education as a Lifelong Process. The German National Educational Panel Study (NEPS).* Hrsg. von Hans-Peter Blossfeld, Hans Günther Roßbach und Jutta von Maurice, S. 5–17. DOI: 10.1007/s11618-011-0178-3.

Blossfeld, Hans-Peter, Hans Günther Roßbach und Jutta von Maurice, Hrsg. (2011). *Zeitschrift für Erziehungswissenschaft* Special Issue 14: *Education as a Lifelong Process. The German National Educational Panel Study (NEPS).*

Leopold, Thomas, Marcel Raab und Jan Skopek (2011). *Starting Cohort 6: Adults (SC6). SUF-Version 1.0.0. Data Manual.* Techn. Ber. Bamberg: National Educational Panel Study (NEPS), University of Bamberg.

Wenzig, Knut (2012). *NEPS-Daten mit DOIs referenzieren.* RatSWD Working Paper Series. 202. Rat für Sozial- und Wirtschaftsdaten, Berlin.

A Anhang – ERD

Die Metadaten sind in vielen Tabellen einer SQL-Datenbank strukturiert abgelegt. Im grünen Bereich finden sich Informationen über Variablen und Schemata, gelb unterlegt sind die Tabellen mit Informationen über Items, rosa Informationen über Instrumente (also Erhebungsinstrumente und SUFs) und im blauen Bereich finden sich Tabellen über das NEPS-Studiendesign.

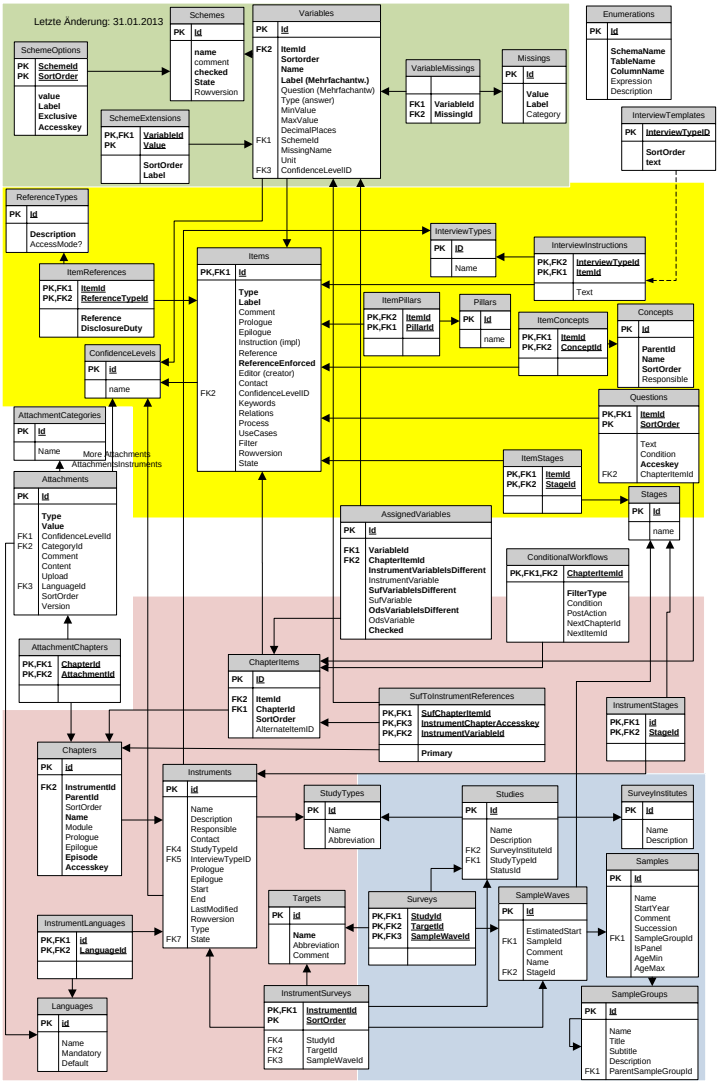

Abbildung 10: Tabellen in der SQL-Metadatenbank (Autor: Gerhard Kraft)