Institut für Deutsche Sprache

Sylvia Dickgießer

unter Mitarbeit von Joachim Gasch

Metadatenschemata in der Datenbank für Gesprochenes Deutsch (DGD 2.0)

Stand: 2011-07-01

Inhaltsverzeichnis

Vorbemerkung

Eine erste Beschreibung der Metadatenschemata in der Datenbank für Gesprochenes Deutsch (DGD 2.0), die unter Mitarbeit von Caren Brinckmann und Joachim Gasch entstanden war, wurde im September 2009 vorgelegt. Die 2009 beschriebenen Schemata wurden seitdem an zahlreichen IDS-Korpora erprobt, zuletzt am Forschungs- und Lehrkorpus gesprochenes Deutsch (FOLK). Im Zuge dieser Arbeiten sind wir auf einige problematische Stellen in den ersten Fassungen aufmerksam geworden, die wir verbessern konnten. Nun möchten wir über den aktuellen Stand der Entwicklung informieren.

Wir danken allen Kolleginnen und Kollegen in der Abteilung Pragmatik, die uns bei der Weiterentwicklung der Schemata unterstützt haben. Besonders hervorheben möchten wir die sorgfältige und geduldige Arbeit von Bistra Angelova und Kunduz Duyshenova an der Dokumentation des Korpus „Deutsch heute“ (DH). Sie haben uns auf praktische Probleme aufmerksam gemacht und wiederholte Änderungen der DH-spezifischen Schemata mit Verständnis und freundlicher Nachsicht ertragen.

1. Einleitung

Die Datenbank für Gesprochenes Deutsch (DGD 2.0) enthält eine Metadatenkomponente, die auf einem neuen Modell für die Dokumentation von Korpora der gesprochenen Sprache beruht und vier darauf aufbauende (XML-)Schemata umfasst.

Die Entwicklung dieser Komponente orientierte sich an folgenden Richtlinien:

- Unabhängigkeit von spezifischen Forschungsansätzen
- Vermittlung zwischen projektübergreifenden und projektspezifischen Anforderungen
- detaillierte Datenstruktur
- kalkulierte Redundanz
- validierbare Datenerfassung
- konsistente zentrale Datenspeicherung
- variable benutzerfreundliche Darstellung
- effektives korpusübergreifendes Retrieval
- datenschutz- und datensicherheitsgerechte Benutzerverwaltung

Die Unabhängigkeit von spezifischen Forschungsansätzen soll es uns ermöglichen, Daten aus verschiedenen Bereichen in übergeordnete Strukturen integrieren zu können.

2. Externe Metadatenschemata

Die Anzahl empfohlener Metadatenschemata ist beachtlich. Die bekanntesten sind Dublin Core (DC) [1], die Systematik des Open Language Archives (OLAC) [2], die der Text Encoding Initiative (TEI) [3], MPEG 7 [4] und schließlich die Schemata der ISLE Meta Data Initiative (IMDI) [5].

Bei der Auswahl und Bewertung möglicher Vorgaben für unsere Arbeit stützten wir uns v.a. auf eine Studie von Thorsten Trippel und Tanja Baumann, die auf der Suche nach einem geeigneten Metadatenstandard für „multimodale Korpora“ („linguistische Korpora“) die Schemata von DC, OLAC, TEI und IMDI miteinander verglichen und auf ihre Nützlichkeit für die Dokumentation solcher Korpora hin überprüft haben. Die Autoren kamen zu dem Ergebnis:

„Für die Archivierung von Ressourcen sind verschiedene Standards definiert und betrachtet worden: Dublin Core: kleinster gemeinsamer Nenner von Metadaten […], wobei der Schwerpunkt auf der Katalogisierung von Ressourcen liegt. OLAC: DC Erweiterung für mehrsprachige und vor allem auch in anderen Medien vorliegenden Ressourcen. TEI: Struktur für Metadaten für gedruckte und textuelle Medien […], wobei weder Mehrsprachigkeit noch andere Medien vorgesehen sind. IMDI: geeignetster Standard, da er die anderen Standards konzeptuell einschließt und gleichzeitig Probleme mehrsprachiger Ressourcen und verschiedener Medien berücksichtigt. Einzig die fehlenden Datenkategorien auf Annotationsebene stellen ein Problem dar, wobei aber auch in anderen Standards hierfür keine Kategorien bekannt sind.“ [6]

IMDI hat bislang zwei Schemata für die Dokumentation linguistischer Ressourcen bereitgestellt: Ein Schema für eine Korpusbeschreibung (catalogue descriptions) und eines für die Dokumentation von Korpusbestandteilen (session descriptions). Diese Schemata wurden für das Retrieval (mit Hilfe spezieller Browser) und für die Publikation von Metadaten konzipiert. Sie werden u.a. im Max-Planck-Institut für Psycholinguistik in Nijmegen verwendet.

„Session“ wird folgendermaßen erläutert: „The session concept bundles all information about the circumstances and conditions of the linguistic event, groups the resources belonging to this linguistic event, records the administrative information of the event and describes the content of the event. Since version 3.0. also written resources other than annotations can be included in a session. For written resources the definition of session is extended to include all documents that pertain to the creation, analysis and commentary of a document.” [7]

Wir haben das IMDI-Session-Schema geprüft und sind zunächst auf zwei für uns problematische Punkte aufmerksam geworden:

a) Um eine Vergleichbarkeit der Daten gewährleisten und zugleich sehr heterogenen Datenbeständen gerecht werden zu können, enthält das Schema nur eine relativ kleine Anzahl verbindlicher Informationselemente. Daneben sind in allen Abschnitten optionale „description elements“, in denen unstrukturierte Texte abgelegt werden können, sowie optionale „keys“ vorgesehen, die es verschiedenen Forschungsgruppen ermöglichen, gruppenspezifische Strukturen in das Schema zu integrieren. Dadurch wird eine große Flexibilität gewährleistet, aber auch eine Beliebigkeit gefördert, die für unsere Zwecke nicht sinnvoll ist.

b) Das Session-Konzept bezieht sich auf „linguistic events“ und speichert alle Daten über die sozialen Kontexte („circumstances and conditions“) dieser „linguistic events“ sowie alle Daten für die an einer „Session“ Beteiligten („actors“), in einem Schema. Das führt zu Redundanzen in der Datenbasis, wenn mehrere „linguistic events“ in einem sozialen Kontext zu beschreiben sind und wenn einzelne Personen an mehreren dieser „linguistic events“ beteiligt waren.

In Anbetracht dieser Besonderheiten entschieden wir uns für die Entwicklung eines eigenen Modells unter Berücksichtigung vorhandener Schemata (wie IMDI) und Empfehlungen für die Dokumentation von Korpora der gesprochenen Sprache, wie z.B. die des Bayerischen Archivs für Sprachsignale (BAS) [8].

3. IDS-Datenmodell für die Dokumentation von Korpora der gesprochenen Sprache

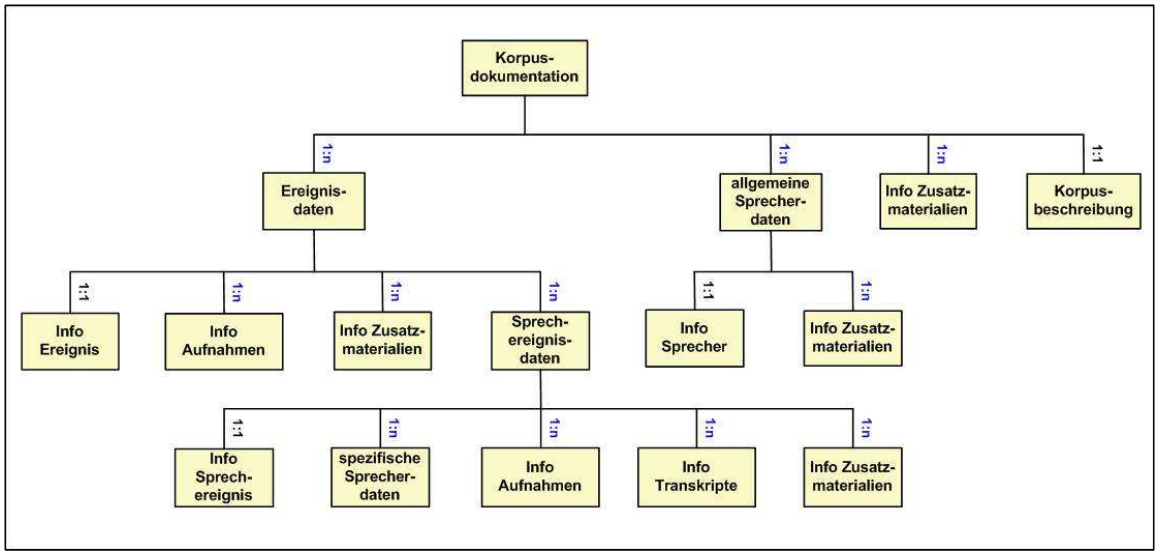

Abb. 1, Datenmodell für die Korpusdokumentation

Abb. 1 zeigt unser Datenmodell für die Dokumentation von Korpora der gesprochenen Sprache, das vier Bereiche vorsieht, die mithilfe von (XML-)Schemata strukturiert werden: einen Bereich für Ereignisdaten, einen Bereich für ereignisübergreifende, allgemeine Sprecherdaten, einen Block für Informationen über Zusatzmaterialien auf Korpusebene (z.B. Transkriptionskonventionen oder Texte, die von allen Informanten vorgelesen wurden) und eine Korpusbeschreibung.

Dieses Modell hebt sich von dem Modell, das der IMDI-Session-Beschreibung zugrundeliegt, v.a. dadurch ab, dass unterschieden wird zwischen Ereignis und Sprechereignis und dass Sprecherdaten in zwei Bereichen abgelegt werden. Damit möchten wir übermäßige Redundanzen in der Datenbasis vermeiden.

Im Laufe der Arbeiten an den Schemata haben wir ein neues System von Kennungen entwickelt. Kennungen sind systematische, eindeutige Kurzbezeichnungen für die Bezugsobjekte der Dokumentation: Ereignis, Sprechereignis, Sprecher, verschiedene Korpusbestandteile und deren technische Fassungen. Diese aufeinander abgestimmten Kurzbezeichnungen werden im Zusammenhang mit den Schemata in den Abschnitten 5. bis 7. vorgestellt.

4. Generische Schemata und projektspezifische Subschemata

Um zwischen projektübergreifenden und projektspezifischen Anforderungen vermitteln zu können, wurden zunächst umfangreiche Sammlungen von Informationselementen für die Bereiche Ereignisdaten und Sprecherdaten zusammengestellt, strukturiert und in (XML-)Schemata übertragen.

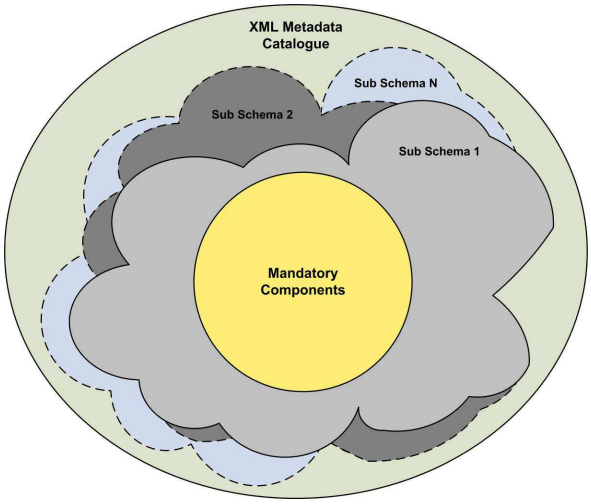

Abb. 2, Generisches Schema (Catalogue) und projektspezifische Subschemata

Mit diesen generischen Schemata werden Standards gesetzt. Sie enthalten obligatorische und fakultative Komponenten, Felddefinitionen und Standardwerte und bilden die Grundlage für projektspezifische Subschemata. Bei der Ableitung eines Subschemas aus dem generischen Schema müssen zunächst alle obligatorischen Komponenten übernommen werden. Diese werden ergänzt durch eine Auswahl fakultativer Komponenten, die für das auswählende Projekt verbindlich werden. Darüber hinaus können einzelne, in den generischen Schemata vorgegebene Werte an Projektbedürfnisse angepasst werden. Diese Anpassung geschieht durch die Spezifikation projektspezifischer Muster, mit denen die eingegebenen Werte schon bei der Erfassung verglichen werden, und eine Vorbelegung von Feldern mit projektspezifischen Werten, u.a. in Form von Auswahllisten, wobei die Vorgaben verschiedener Projekte koordiniert werden sollten.

5. Generisches Schema für die Dokumentation von Korpusbestandteilen auf der Ereignisebene

Die Kategorie „Ereignis“ dient als Startknoten eines generischen (XML-)Schemas, das folgende Informationen vorsieht: Angaben über Aufzeichnungsobjekte (Ereignis, Sprechereignis, Sprecher), Angaben über Korpusbestandteile (Audioaufnahmen, Videoaufnahmen, Transkripte, Zusatzmaterialien auf Ereignis- und Sprechereignisebene) sowie eine Dokumentationsgeschichte.

Das Schema enthält obligatorische und fakultative Komponenten. Obligatorische Komponenten sind in allen projektspezifischen Subschemata zu berücksichtigen, fakultative Komponenten

stehen zur Wahl. Wenn Sie verwendet werden, müssen alle Kennungsfelder und alle Felder, die ein Fragezeichen (?) enthalten, bearbeitet werden.

Eingaben für fehlende Daten in Feldern mit Fragezeichen (?) sind standardisiert: „Nicht dokumentiert“ bedeutet: Es kann ein Datum geben, das bei der Datenerfassung jedoch nicht bekannt ist. Ein Beispiel dafür wäre: „Sonstige_Bezeichnungen: Nicht dokumentiert“ - zu lesen als: „Im Projekt kann eine andere Kurzbezeichnung als die Ereigniskennung vergeben worden sein, die jedoch nicht bekannt ist.“ „Nicht vorhanden“ bedeutet: Es gibt kein Datum. Ein Beispiel dafür wäre: „An_E_teilnehmende_Techniker: Nicht vorhanden“ - zu lesen als: „An diesem Ereignis hat kein Techniker teilgenommen.“ In einem Feld („Datenrate“) kann der Wert „Nicht relevant“ verwendet werden.

Das an vielen Stellen vorgesehene Feld „Anmerkungen“ ist für Anmerkungen zu Angaben in anderen Feldern und für nicht kategorisierte Angaben vorgesehen. Das Feld kann leer bleiben.

Einzelne Komponenten des Schemas wurden als iterativ gekennzeichnet, d.h. dass sie bei der Datenerfassung vervielfältigt werden können.

Die nachfolgenden Abbildungen stammen aus einem projektneutralen Erfassungsformular, das zu Demonstrationszwecken angelegt wurde.

5.1. Ereignis

Unter „Ereignis“ (E) verstehen wir eine Phase des sozialen Geschehens, die von Beteiligten bzw. Korpusproduzenten als abgrenzbare Einheit wahrgenommen und aufgezeichnet wird. Diese Definition ist aus arbeitspraktischen Gründen bewusst sehr allgemein gehalten. Wir stellen lediglich ein für die Dokumentation von Korpusbestandteilen relevantes Konzept bereit, keine linguistischen Segmentierungskriterien. Zur Veranschaulichung unseres Ereigniskonzeptes nennen wir im Folgenden drei Beispiele:

Im Korpusprojekt „Deutsch heute“ (DH) gelten mehrstündige Aufnahmesitzungen in Schulen und Volkshochschulen, die von Projektmitarbeitern geleitet wurden, als zu dokumentierende Ereignisse. Das IDS-Korpus „Stadtsprache: Mannheim“ enthält u.a. Aufzeichnungen von Treffen sozialer Gruppen in bestimmten Stadtteilen. Jedes dieser Gruppentreffen kann als Ereignis dokumentiert werden. Ein im IDS-Korpus „Biographische und Reiseerzählungen“ aufgezeichnetes Ereignis wurde folgendermaßen beschrieben: „Gemeinsames Kaffeetrinken während einer Seminarpause. Das Treffen zwischen Studentinnen und Dozentinnen wurde organisiert, um Reiseerzählungen aufzuzeichnen.“

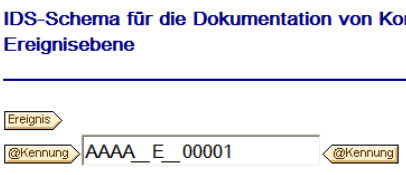

Abb. 3, Ereignis - Kennung

An erster Stelle der Ereignisdaten wird eine Kennung eingetragen, die eine vierstellige Korpuskennung, den Kennbuchstaben E (für „Ereignis“) und eine fünfstellige laufende Nummer umfasst. Ein Beispiel für eine Ereigniskennung finden Sie in Abb. 3.

5.1.1. Basisdaten

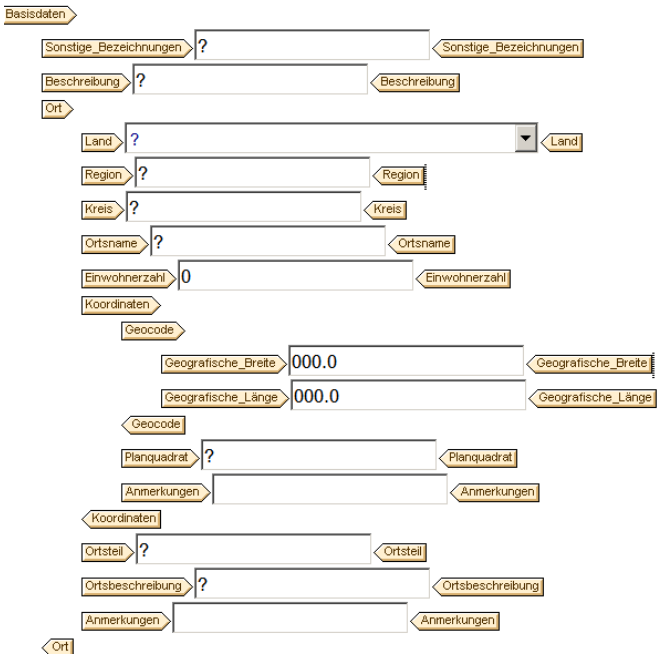

Abb. 4, Ereignis - Basisdaten (1)

Die Ereignis-Basisdaten werden mit dem Feld „Sonstige_Bezeichnungen“ eröffnet. Damit sind projektinterne Kurzbezeichnungen des Ereignisses gemeint. Im IDS-Projekt „Deutsch heute“ z.B. wurden dreistellige Ortskürzel wie „ODF“ (für „Oberstdorf“) benutzt.

Im Anschluss daran wird eine kurze Charakterisierung des aufgezeichneten Ereignisses erwartet. Hier sollte man nach Möglichkeit auch angeben, ob das Ereignis geplant oder nicht geplant war sowie ob und wann die Beteiligten über die Aufnahmen informiert wurden. Unsere erste Beispielbeschreibung stammt aus der Dokumentation des Korpus „Deutsch heute, die zweite aus der Dokumentation des Korpus Grundstrukturen: Freiburger Korpus: 1.) „Geplante Aufnahmeaktion im Rahmen des Spracherhebungsprojekts Deutsch heute, wobei von jedem Sprecher das gleiche Material erhoben wird (Lesesprache, Interview, Maptask). Die Sprecher waren im Vorfeld über die Aufzeichnungen informiert worden.“ 2.) „Lesung von Günther Grass aus "Katz und Maus" und anschließende Diskussion.„

Im Komplex „Ort“ sind Angaben über den Ort zusammengefasst, an dem das jeweilige Ereignis stattfand. Für Werte im Feld „Land“ gibt es eine ISO-Liste. Das Feld „Region“ ist für die amtliche Bezeichnung eines Bundeslandes, eines Kantons oder einer Provinz vorgesehen. In die Felder „Kreis“ und „Ortsname“ sollen ebenfalls amtliche Bezeichnungen eingetragen werden. Die Komponente „Koordinaten“ ist fakultativ. Wenn sie von einem Projekt gewählt wird, ist entweder der Geocode oder das Planquadrat (Kategorie im DSAv-Katalog [9]) des Ortes zu verzeichnen. Im Feld „Ortsteil“ kann der Name des Ortsteils, in dem das Ereignis stattfand, notiert werden. Weitere Informationen über den Ort kann man im Feld „Ortsbeschreibung“ festhalten.

Im Anschluss an die Ortsangaben soll der gesellschaftliche Kontext eines Ereignisses charakterisiert werden. Dafür ist das Feld „Institution“ vorgesehen. „Institution“ verstehen wir i.S.v. „Organisation“. Im Feld „Räumlichkeiten“ kann man Angaben zur räumlichen Umgebung des Ereignisses notieren.

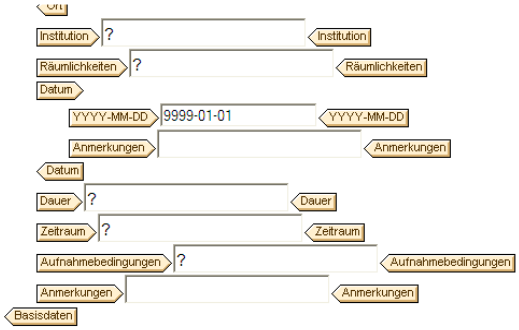

Abb. 5, Ereignis - Basisdaten (2)

Das Modul „Datum“ ist für Angaben über das Datum, an dem das Ereignis stattfand, vorgesehen. Wenn sich ein Ereignis über mehrere Tage erstreckte, sollte nur das Anfangsdatum aufgenommen werden. Das zugehörige Feld „Anmerkungen“ bietet die Möglichkeit, auf ungenaue Daten hinzuweisen. Die Dauer des aufgezeichneten Ereignisses ist im gleichnamigen Feld zu erfassen. An dieser Stelle sind auch ungenaue Angaben wie z.B. „Ca. 6 Stunden“ möglich. Über den Zeitraum, innerhalb dessen ein Ereignis stattfand, kann im gleichnamigen Feld z.B. folgendermaßen informiert werden: „Von 16 bis 17 Uhr“, „Vormittags“, „Zwei Tage“. Ereignisse, die zu völlig unterschiedlichen Zeiten stattfanden, gelten als unterschiedliche Ereignisse, die in verschiedenen Dokumenten zu beschreiben sind.

Unter dem Stichwort „Aufnahmebedingungen“ sollten Angaben über besondere Umstände der Aufzeichnungsaktion, z.B. Zeitdruck oder problematische akustische Verhältnisse, erfasst werden.

5.1.2. Rundfunksendung

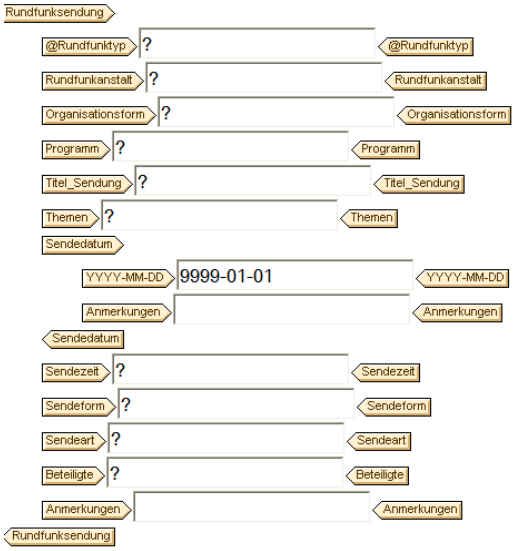

Abb. 6, Rundfunksendung

Übertragungen eines Ereignisses im Rundfunk können im fakultativen Komplex „Rundfunksendung“ dokumentiert werden. Für den Fall, dass Mitschnitte von mehreren Übertragungen zu dokumentieren sind, wurde der Komplex im Schema als iterativ gekennzeichnet.

An erster Stelle ist der Rundfunktyp - Hörfunk oder Fernsehen - anzugeben. Es folgen Felder für den Namen der Rundfunkanstalt, die eine Sendung verantwortet (z.B. „Bayerischer Rundfunk“, „Hessischer Rundfunk“, „Österreichischer Rundfunk“, „ZDF“) und für eine Information

über deren Organisationsform (z.B. „Öffentlich-rechtlich“ oder „Privat“). Im Feld „Programm“ soll der „Kanal“ verzeichnet werden. Mögliche Werte sind z.B. „Kinderkanal“, „SWR-1“, „WDR-1-regional“, „RBB-Berlin“. [10]

Für den gesamten Titelkomplex einer Sendung steht das Feld „Titel_Sendung“ bereit. Hier können Haupt- und Untertitel von einzelnen Sendungen und Sendereihen eingetragen werden. Soweit Angaben über das Thema einer Sendung nicht im Titel enthalten sind, kann man sie im Feld „Themen“ verzeichnen.

Die Komponente „Sendedatum“ ist für Angaben über das Datum, an dem das Ereignis übertragen wurde, vorgesehen. Es folgt ein Feld für die Sendezeit, wo die Anfangs- und die Endzeit eingetragen werden sollten. Im Feld „Sendeform“ können Bezeichnungen wie „Bericht“, „Diskussion“, „Interview“, „Magazin“, „Nachrichten“, „Show“ etc. erfasst werden. Das Feld „Sendeart“ ist für Hinweise wie z.B. „Erstausstrahlung“ oder „1. Wiederholung“ etc. gedacht.

Für die Namen von Produzenten (dazu zählen u.a. Regisseure und Moderatoren) und sonstigen Mitwirkenden (z.B. Gäste bei einer Talkshow) wurde das Feld „Beteiligte“ eingerichtet.

5.2. Projekt

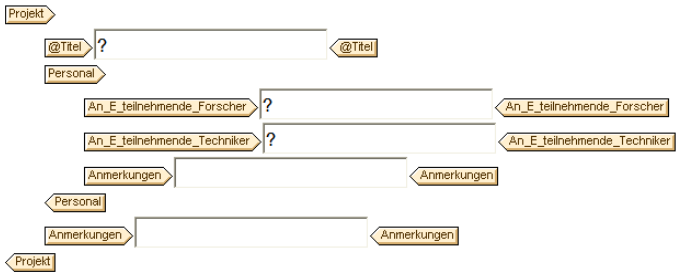

Abb. 7, Ereignis - Projekt

In der Komponente „Projekt“ werden Angaben über das Projekt, in dem das Ereignis aufgezeichnet wurde, zusammengefasst. Um Kooperationen zwischen mehreren Projekten gerecht werden zu können, wurde die Komponente als iterativ gekennzeichnet.

5.3. Quellaufnahme

Unter „Quellaufnahmen“ verstehen wir Rohdaten, Originalaufnahmen von Ereignissen oder Aufnahmen, die für die dokumentierende Stelle Originalcharakter haben. Diese Aufnahmen können Quellen für sprechereignisspezifische Kopien sein. Da ein Korpus nicht unbedingt Originalaufnahmen umfassen muss, ist dieser Komplex fakultativ. Um mehrere Quellaufnahmen dokumentieren zu können, haben wir ihn im Schema als iterativ gekennzeichnet.

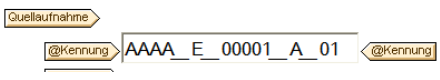

Abb. 8, Quellaufnahme - Kennung

An erster Stelle der Dokumentation einer Quellaufnahme steht eine Kennung, die sich zusammensetzt aus der Kennung des Ereignisses, dem Kennbuchstaben A (für „Aufnahme“) und einer zweistelligen laufenden Nummer. Ein Beispiel finden Sie in Abb. 8.

5.3.1. Basisdaten

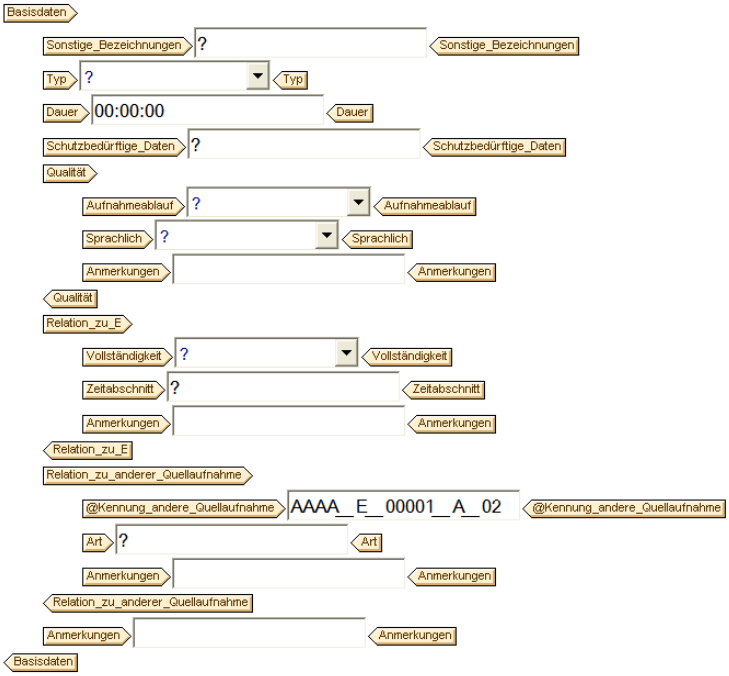

Abb. 9, Quellaufnahme - Basisdaten

Im ersten Feld der Quellaufnahme-Basisdaten werden „Sonstige_Bezeichnungen“ abgefragt. Damit sind eventuell im Projekt vergebene Kurzbezeichnungen gemeint. Für eine Aufnahme aus dem Korpus „Stadtsprache: Mannheim“ z.B. könnte man an dieser Stelle eine im Projekt gewählte „Diskursnummer“ (z.B. „2036.50“) eintragen.

Quellaufnahmen können unterschiedlichen Typs sein: Audioaufnahme, Videoaufnahme und ggf. auch Tonkopie von Videoaufnahme. Für Angaben über die Dauer der jeweiligen Aufnahme wurde ein Zeitfeld mit dem Format hh:mm:ss vorbereitet.

Quellaufnahmen können Daten enthalten, die nach dem Willen der Urheber und aus datenschutzrechtlichen Gründen Außenstehenden nicht kenntlich werden dürfen, wie z.B. persönliche Sprecherdaten. Für Informationen über solche Daten ist das Feld „Schutzbedürftige_Daten“ vorgesehen.

Die Komponente „Qualität“ ist fakultativ. Wir haben die Felder „Aufnahmeablauf“ und „Sprachlich“ v.a. im Hinblick auf eine mögliche Übernahme der DSAv-Katalogdaten in die Struktur eingesetzt. Auf Seite 5 des DSAv-Gesamtkatalogs [9] sind Kriterien zusammengestellt, die zu einer schlechten Beurteilung katalogisierter Aufnahmen führten: a) Sprachliche Qualität - „Zahnschäden, Nuscheln, Lispeln, Mund- oder Prothesengeräusche, Stottern, scharfe, zischende s oder z, Asthma, Heiserkeit, […] starke Befangenheit oder Erregung, zu leise oder zu laute Sprache, […]“; b) Aufnahmeablauf - „Pausen, Zwischenfragen, Absätze, stockendes Erzählen, Dazwischenreden.“

Unter der Überschrift „Relation_zu_E“ werden Informationen über das Verhältnis der jeweiligen Quellaufnahme zum Ereignis erfasst. Im Feld „Vollständigkeit“ wird eingetragen, ob eine vollständige oder eine unvollständige Aufnahme eines Ereignisses dokumentiert wird. Die Angabe „Unvollständig“ sollte im Feld „Zeitabschnitt“ präzisiert werden. „Zeitabschnitt“ meint den in der jeweiligen Quellaufnahme aufgezeichneten Zeitabschnitt des Ereignisses. Wenn die genaue Zeit nicht zu ermitteln ist, kann hier darüber informiert werden, welcher Abschnitt des Ereignisses in der jeweiligen Aufnahme festgehalten ist, z.B. „1. Abschnitt“, „2. Abschnitt“ usw. Bei vollständigen Aufnahmen wird der Wert „Vollständig“ erwartet.

Die Komponente „Relation_zu_anderer_Quellaufnahme“ ist fakultativ und iterativ. Wenn es nur eine Quellaufnahme gibt, ist sie nicht relevant. An dieser Stelle kann man z.B. auf Überlappungen von Aufnahmen hinweisen. Benötigt werden die Kennung der anderen Quellaufnahme und eine Information über die Art der Beziehung.

5.3.2. Aufnahmetechnik

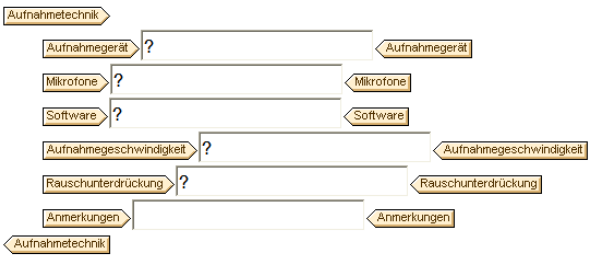

Abb. 10, Quellaufnahme – Aufnahmetechnik

Unter dem Stichwort „Aufnahmetechnik“ werden Informationen über die Aufnahmeapparatur (Aufnahmegerät, Mikrofone), eine ggf. eingesetzte Aufzeichnungssoftware, die Aufnahmegeschwindigkeit (bei Spulentonbandaufnahmen relevante Angabe in cm/s) und Rauschunterdrückungsverfahren (z.B. Dolby B) zusammengefasst.

5.3.3. Technische Fassungen

Quellaufnahmen liegen in bestimmten technischen Fassungen vor. Das können analoge und / oder digitale Fassungen sein. Für jeden Typ gibt es einen eigenen Abschnitt im Schema. Beide Abschnitte sind fakultativ und iterativ, wenigsten ein Abschnitt muss bei der Erstellung projektspezifischer Schemata übernommen werden.

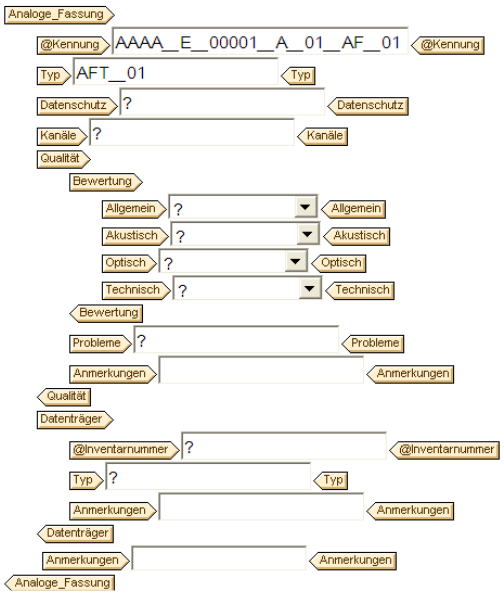

Abb. 11, Quellaufnahme - Analoge Fassung

Für jede technische Fassung wird eine Kennung generiert. Diese Kennung ist zusammengesetzt aus der Kennung der jeweiligen Quellaufnahme, dem Kürzel AF (für „Analoge_Fassung“) bzw. DF (für „Digitale_Fassung“) und einer zweistelligen laufenden Nummer.

Im Anschluss an die Kennungen werden Typen analoger und digitaler Fassungen benannt. Grundlage für eine Typisierung analoger Fassungen sind Datenschutz und Kanäle, für die Typisierung digitaler Fassungen sind außerdem noch technische Daten (z.B. das Dateiformat) relevant. Als Typenbezeichnungen dienen die Kürzel AFT (analoge Fassung) und DFT (digitale Fassung) in Verbindung mit einer zweistelligen Nummer. Beispiele für Kennungen und Typenbezeichnungen finden Sie in den folgenden Abbildungen.

„Datenschutz“ meint technische Maßnahmen zum Datenschutz, wie z.B. die Überlagerung von Personennamen in Aufnahmen. Das Feld „Kanäle“ steht für Angaben wie „Mono“ oder „Stereo“ bereit.

Die Komponente „Bewertung“ im Abschnitt „Qualität“ ist fakultativ. Die Bewertung unterschiedlicher Aspekte der Aufnahmequalität soll anhand einer Skala erfolgen, die in Abb. 12 dargestellt wird.

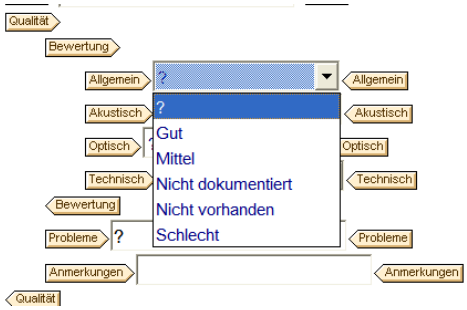

Abb. 12, Aufnahmequalität - Bewertungsskala

Das Feld „Allgemein“ soll eine Bewertung des Gesamteindrucks aufnehmen. Da uns für die anderen Aspekte kein neuer Kriterienkatalog vorliegt, verwenden wir zur Veranschaulichung ältere Hinweise aus dem DSAv-Gesamtkatalog [9] (S. 5).

Die akustische Aufnahmequalität können lt. Katalog beeinträchtigen: „halliger Aufnahmeraum, Unruhe bzw. Störgeräusche im Sprecherraum oder von außen, Sprecher zu nahe am Mikrofon (harte p, t, k), wechselnder bzw. ungünstiger Mikrophonabstand [sic!] zum Sprecher“. Folgende Erscheinungen führen lt. Katalog zu einer negativen Bewertung der technischen Aufnahmequalität: „Netzbrummen, Leitungston, Kopiereffekt, verzerrte Modulation durch elektrische Fehler an den Geräten, Bandfehler“. In dieser älteren Aufzählung fehlt das bekannte Phänomen „Verzerrung“, das durch Übersteuerung bei analogen Aufnahmen (und in neuerer Zeit auch durch Fehler bei der Digitalisierung) entstehen kann.

Das Feld „Optisch“ ist für eine Bewertung der optischen Qualität von Videoaufnahmen vorgesehen. Im Feld „Probleme“ können konkrete Qualitätsprobleme benannt werden. Mögliche Werte sind z.B.: „Lauter Verkehrslärm“ ; „Leises Brummen“ ; „Die Aufnahme ist stellenweise verzerrt“ ; „Die Aufnahme ist an vielen Stellen unverständlich, da oft durcheinander geredet wird“.

Da eine technische Fassung auf verschiedenen Datenträgern gespeichert sein kann, haben wir das entsprechende Modul im Schema als iterativ gekennzeichnet. An erster Stelle wird eine eindeutige Inventarnummer des zu dokumentierenden Datenträgers erwartet. Im nächsten Schritt sollte man über den Typ des Datenträgers (z.B. Kompaktkassette oder Tonband) informieren.

Für den Abschnitt „Digitale_Fassung“ sind außer den bisher beschriebenen Komponenten weitere Strukturelemente relevant.

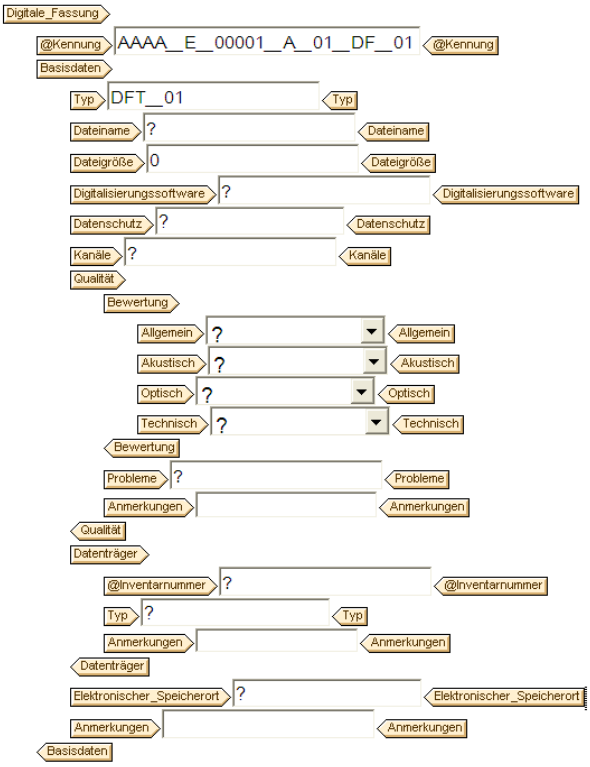

Abb. 13, Quellaufnahmen – Digitale Fassung – Kennung und Basisdaten

An zweiter Stelle im Modul „Basisdaten“ ist der Dateiname zu nennen. Die Dateigröße soll in Bytes angegeben werden. Im Feld „Digitalisierungssoftware“ sollte notiert werden, welche Software ggf. bei der Digitalisierung einer analogen Fassung verwendet wurde. [11] Elektronischer_Speicherort“ verlangt einen Pfadnamen oder eine URL.

Im nächsten Abschnitt unterscheiden wir tontechnische und videotechnische Daten. Für Felder in diesen Modulen bieten wir im Folgenden kurze Erläuterungen an. Weitere Informationen finden Sie u.a. im Gesprächsanalytischen Informationssystem (GAIS) http://gais.ids-mannheim.de/ und unter den unten genannten Adressen.

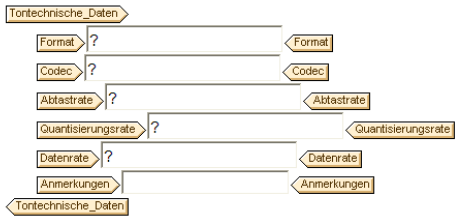

Abb. 14, Quellaufnahme - Digitale Fassung – tontechnische Daten

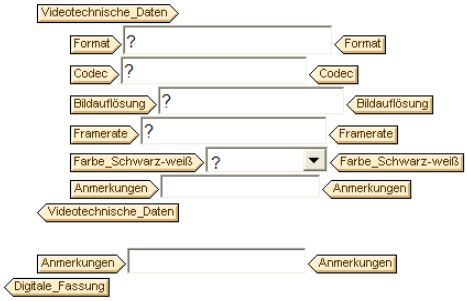

Abb. 15, Quellaufnahme - Digitale Fassung – videotechnische Daten

Im Feld „Format“ werden Angaben über das Dateiformat und das Digitalisierungsverfahren bzw. Audioformat erfasst. Ein möglicher Feldwert wäre: „WAVE (Linear PCM)“. Informationen über relevante Audioformate finden Sie unter der Adresse http://de.wikipedia.org/wiki/Audioformat.

Als „Codec“ bezeichnet man ein Verfahren bzw. Programm, das Daten oder Signale digital kodiert und dekodiert. Unter der Adresse http://de.wikipedia.org/wiki/Codec finden Sie eine Liste mit Namen gängiger Codecs.

„Abtastung“ (engl. sampling) bezeichnet die Registrierung von Messwerten zu diskreten, meist äquidistanten Zeitpunkten. Aus einem zeitkontinuierlichen Signal wird so ein zeitdiskretes Signal gewonnen. Die Anzahl der Abtastungen pro Zeiteinheit wird Abtastrate genannt und meist in Hertz (Hz = Anzahl pro Sekunde) angegeben. Mögliche Werte sind z.B. „44100“ oder „48000“.

Nach der Abtastung erfolgt die Quantisierung des zeitdiskreten, aber noch wertkontinuierlichen Signals. Dadurch entsteht ein zeit- und wertdiskretes Signal. Die Quantisierungsrate (auch Samplingtiefe oder Bittiefe) gibt die Anzahl der Bits an, die bei der Quantisierung pro Abtastwert verwendet werden. Typische Quantisierungsraten sind 8, 16 und 24 Bit.

Bei komprimierten Daten wird die Datenrate relevant - die Anzahl der Informationseinheiten, die pro Zeiteinheit gespeichert werden. Sie wird in kBit/s angegeben.

Die Bildauflösung meint in unserem Kontext die Größe eines Einzelbildes (Breite x Höhe), wobei in der horizontalen Richtung (Breite) die max. Zeilenanzahl und in der vertikalen (Höhe) die max. Spaltenanzahl einer Bilddarstellung angegeben wird. Die Bildauflösung gibt also Auskunft darüber, mit welche max. Qualität (Pixel) ein Bild dargestellt werden kann. Die Bildauflösung ist für bestimmte Formate (VHS, DVD, HD usw.) normiert. Bis zur Einführung vom HD waren in Europa durch das PAL System und den alten Fernsehgeräten mit einem Seitenverhältnis von 4:3 nachfolgende Bildauflösungen gebräuchlich: VHS 400 x 576 (4:3) = 230400 Pixel;
DVD 720 x 576 (4:3 oder 16:9) = 414720 Pixel. Durch die Einführung von HD werden in Zukunft die o.g. Bildauflösungen, Bildformate durch nachfolgende ersetzt: HD 1920 x 1080(p) (16:9) = 2073600 Pixel. [12]

Die Framerate (Bildwiederholrate) z.B. eines Videofilms gibt Auskunft über die Anzahl der Einzelbilder, die in 1 Sekunde Film abgespielt bzw. projiziert werden. Fernsehfilme in Europa werden in der Regel mit 25 Bildern in der Sekunde abgespielt bzw. gesendet (25 fps). Kinofilme werden in der Regel mit 24 Bildern in der Sekunde abgespielt bzw. projiziert (24 fps). Die Bildwiederholrate von 25 Bildern pro Sekunde ist in Europa in der Fernsehnorm festgelegt. In der Regel sollte bei der Erstellung von Videofilmen eine Bildwiederholrate von 20-25 Bildern pro Sekunde nicht unterschritten werden, da sonst das menschliche Auge die Bewegungen im Videomaterial nicht mehr als flüssige kontinuierliche Bewegung wahrnimmt. [12]

5.3.4. Archivierung und Distribution

In den Abb. 16 und 17 werden die Bausteine „Archivierung“ und „Distribution“ vorgestellt, die an verschiedenen Stellen des Ereignisschemas vorkommen. Sie sollen Informationen über rechtliche und organisatorische Aspekte der Korpusbestandteile aufnehmen.

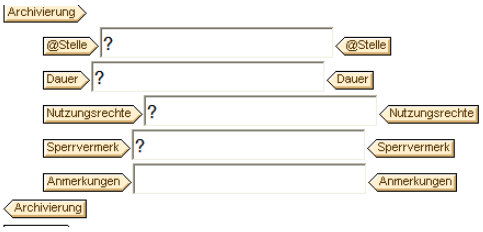

Abb. 16, Quellaufnahme - Archivierung

Das Modul „Archivierung“ wurde im Schema als iterativ gekennzeichnet. Zunächst soll der Name der archivierenden Stelle vermerkt werden. Es folgt ein Feld für Informationen über die vorgesehene Archivierungsdauer. Hier könnte z.B. „Bis 2018“ oder „Langfristig“ stehen. Die Nutzungsrechte der archivierenden Stelle können von der ausschließlichen wissenschaftlichen Auswertung durch den Aufnahmeleiter bis hin zur Veröffentlichung einer Aufnahme im Internet reichen. Sperrvermerke, wie z.B. „Bis 2010 für Externe gesperrt“, können die Nutzungsmöglichkeiten einschränken.

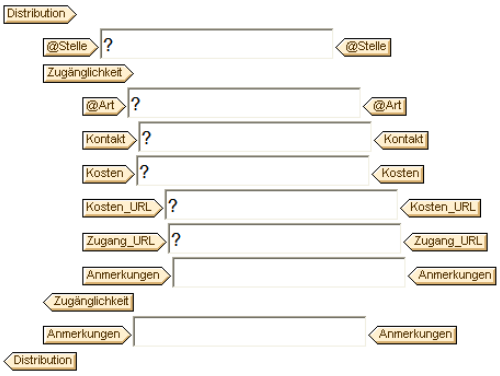

Abb. 17, Quellaufnahme - Distribution

Das Modul „Distribution“ ist ebenfalls iterativ und umfasst neben einem Feld für den Namen der für die Distribution zuständigen Stelle die iterative Komponente „Zugänglichkeit“. In dieser Komponente sollen folgende Angaben verzeichnet werden: Art der Zugänglichkeit, E-Mail-Kontaktadresse, Angaben über die Kosten, ggf. eine URL dieser Angaben sowie ggf. eine URL, die einen direkten Zugang zum jeweiligen Korpusbestandteil ermöglicht.

5.4. Zusatzmaterial

Unter „Zusatzmaterial“ auf der Ereignisebene verstehen wir Dokumente, die zusätzlich zu Quellaufnahmen vorhanden sein können. Das können z.B. Reiseberichte der Aufnahmeleiter sein, Protokolle von Aufnahmesitzungen, Fotos von Aufnahmeorten, Notizen zu einer Sitzordnung etc. Der gesamte Komplex „Zusatzmaterial“ ist fakultativ. Für den Fall, dass mehrere Dokumente zu einem Ereignis zu beschreiben sind, wurde er im Schema als iterativ gekennzeichnet.

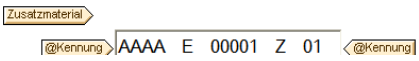

Abb. 18, Zusatzmaterial - Kennung

Die Kennung für Zusatzmaterial auf der Ereignisebene setzt sich zusammen aus der Ereigniskennung, dem Kennbuchstaben Z (für „Zusatzmaterial“) und einer zweistelligen laufenden Nummer. Ein Beispiel finden Sie in Abb. 18.

5.4.1. Basisdaten

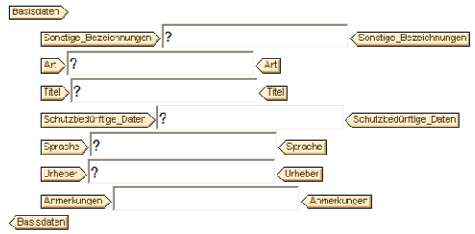

Abb. 19, Ereignis - Zusatzmaterial - Basisdaten

An erster Stelle der Basisdaten für Zusatzmaterial steht das Feld „Sonstige_Bezeichnungen“, wo eine im Erstellungsprojekt für ein Dokument vergebene Kurzbezeichnung abgelegt werden kann. Im Feld „Art“ wird eine Angabe über die Art des Dokuments (z.B. „Skizze der Sitzordnung“) erwartet. Zusatzmaterialien können Daten enthalten, die nach dem Willen der Urheber und aus datenschutzrechtlichen Gründen Außenstehenden nicht kenntlich werden dürfen, wie z.B. persönliche Sprecherdaten, über die man im Feld „Schutzbedürftige_Daten“ informieren kann. Die Sprache, in der ein Textdokument abgefasst ist, sollte ebenso vermerkt werden wie der Urheber eines Dokuments, wobei „Urheber“ für Autoren, Grafiker, Fotografen etc. verwendet wird.

5.4.2. Technische Fassungen

Da zusätzliche Dokumente in verschiedenen technischen Fassungen vorliegen können, haben wir auch an dieser Stelle die fakultativen und iterativen Komponenten „Analoge_Fassung“ und „Digitale_Fassung“ eingefügt. Wenigstens eine Komponente muss bei der Erstellung eines projektspezifischen Schemas gewählt werden.

Zunächst wird eine Kennung abgefragt, die aus der Kennung des Zusatzmaterials, dem Kürzel AF (für „Analoge_Fassung“) bzw. DF (für „Digitale_Fassung“) und einer zweistelligen laufenden Nummer besteht. Beispiele finden Sie in den Abb. 20 und 21.

„Datenschutz“ meint technische Maßnahmen zum Datenschutz, wie z.B. die Maskierung von Personennamen in Texten.

Da eine technische Fassung auf mehreren Datenträgern gespeichert sein kann, wurde dieser Abschnitt im Schema als iterativ gekennzeichnet. An erster Stelle dieses Abschnitts wird eine eindeutige Inventarnummer des zu dokumentierenden Datenträgers erwartet. Im nächsten Schritt ist über den Typ des Datenträgers (z.B. Papier oder Mikrofilm) zu informieren.

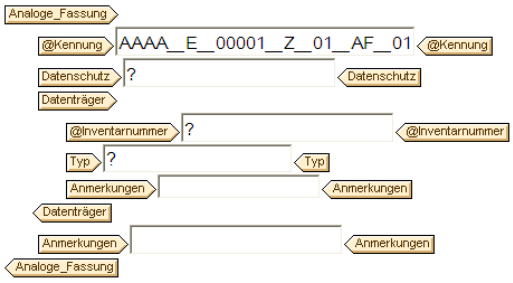

Abb. 20, Ereignis - Zusatzmaterial - Analoge Fassung

Nur für digitale Fassungen relevant sind die Felder „Typ“, „Digitalisierungssoftware“ und „Elektronischer_Speicherort“ sowie die Komponente „Technische_Daten“.

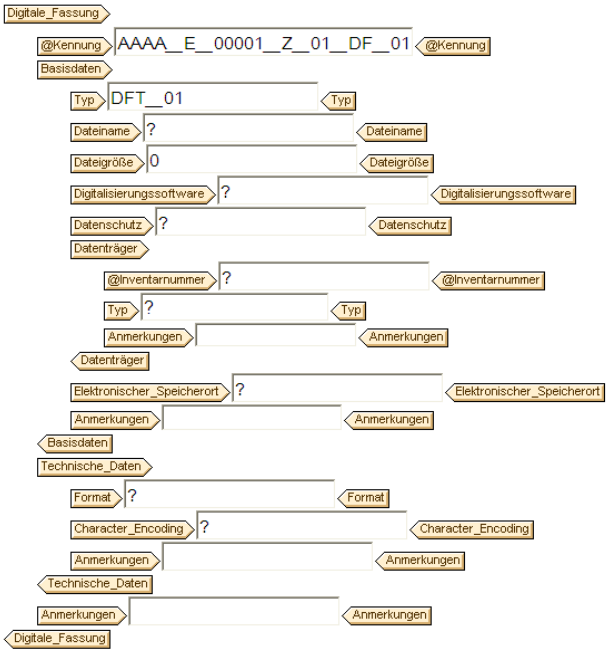

Abb. 21, Ereignis - Zusatzmaterial - Digitale Fassung

Grundlage für eine Typisierung digitaler Fassungen von Zusatzmaterial sind v.a. technische Daten. Als Typenbezeichnungen dient das Kürzel DFT (digitale Fassung) in Verbindung mit einer zweistelligen Nummer.

Die Dateigröße soll in Bytes angegeben werden. Im Feld „Digitalisierungssoftware“ kann das Programm, mit dem eine analoge Fassung digitalisiert wurde, genannt werden. Im Feld „Elektronischer Speicherort“ wird eine URL oder ein Pfadname erwartet.

Das erste Feld der Komponente „Technische_Daten“ ist für eine Information über das Dateiformat vorgesehen. „Character_Encoding“ steht für die Zeichencodierung in einer Textdatei (z.B. ASCII oder UTF-16BE).

5.4.3. Archivierung und Distribution

Die Module „Archivierung“ und „Distribution“, die Informationen über rechtliche und organisatorische Aspekte der Korpusbestandteile aufnehmen sollen, wurden bereits vorgestellt. Die Erläuterungen in Abschnitt 5.3.4. gelten auch für Zusatzmaterial.

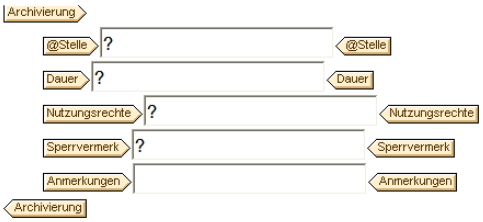

Abb. 22, Ereignis - Zusatzmaterial - Archivierung

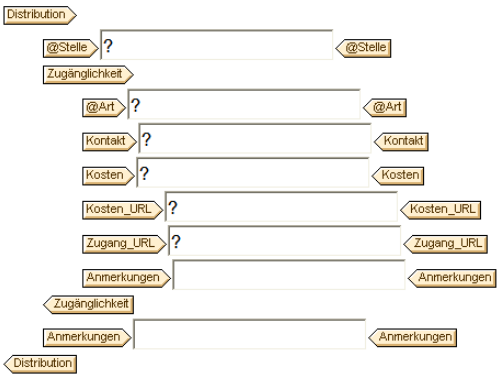

Abb. 23, Ereignis - Zusatzmaterial - Distribution

5.5. Sprechereignis

Unter „Sprechereignis“ (SE) verstehen wir den aufgezeichneten kommunikativen Gehalt eines Ereignisses bzw. Segmente dieses Gehalts. Hier wie in Abschnitt 5.1. gilt: Diese Definition ist aus arbeitspraktischen Gründen bewusst sehr allgemein gehalten. Wir stellen lediglich ein für die Dokumentation von Korpusbestandteilen relevantes Konzept bereit, keine linguistischen Segmentierungskriterien. Zur Veranschaulichung unseres Sprechereigniskonzeptes nennen wir im Folgenden einige Beispiele:

Im Korpusprojekt „Deutsch heute“ gilt jede Aufgabe, die im Rahmen einer mehrstündigen Aufnahmesitzung bearbeitet wurde, als Sprechereignis. Zu diesen Aufgaben gehören u.a. Bildbenennung, Verlesen einer Wortliste, Übersetzung und Interview. Im IDS-Korpus „Stadtsprache: Mannheim“ sind Aufnahmen von Gruppentreffen enthalten, bei denen z.B. Witze erzählt, Klatsch ausgetauscht und gemeinsame Unternehmungen geplant wurden. Solche kommunikativen Sequenzen können als einzelne Sprechereignisse dokumentiert werden. Das IDS-Korpus „Elizitierte Konfliktgespräche“ enthält Aufzeichnungen von Settings, in denen jeweils eine Mutter-Tochter-Dyade zwei Konfliktgespräche führte. Das Thema des ersten Gesprächs wurde von der Mutter eingebracht, das Thema des zweiten von der Tochter. Wir betrachten jedes dieser Gespräche als ein Sprechereignis.

Damit mehrere Sprechereignisse pro Ereignis dokumentiert werden können, wurde dieser Bereich im Schema als iterativ gekennzeichnet.

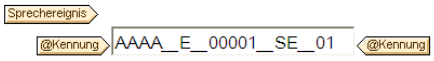

Abb. 24, Sprechereignis - Kennung

An erster Stelle der SE-Daten steht eine Kennung, die die jeweilige Ereigniskennung, das Kürzel SE (für „Sprechereignis“) und eine zweistellige laufende Nummer umfasst. Ein Beispiel finden Sie in Abb. 24.

5.5.1. Basisdaten

Zu Beginn der Basisdaten steht das Feld „Sonstige_Bezeichnungen“, wo im Projekt vergebene Kurzbezeichnungen abgelegt werden können. Im Projekt kann auch ein Titel für das zu dokumentierende Sprechereignis vergeben worden sein, der im gleichnamigen Feld zu erfassen ist.

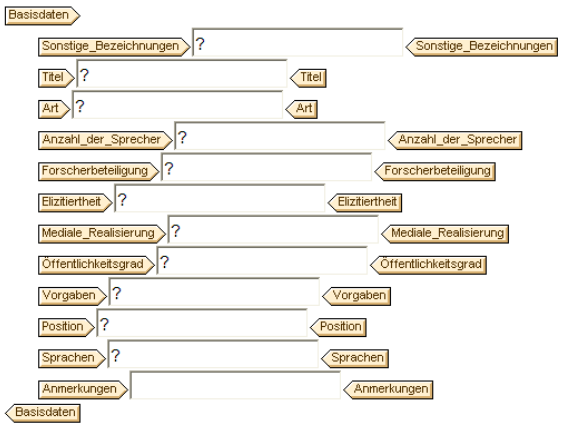

Abb. 25, Sprechereignis- Basisdaten,

Wir verwenden „Art“ anstelle von Kategorien wie „Textsorte“, „Texttyp“, „Interaktionstyp“, „Gesprächstyp“, „Diskurstyp“, „Genre“, „Gattung“, die aus verschiedenen Forschungsansätzen stammen, um Daten aus allen Bereichen aufnehmen zu können. Nach unserer Vorstellung können in dieses Feld mehrere Angaben eingetragen werden. Wir denken dabei an Werte wie „Erzählung“, „Rede“, „Anleitung“, „Beschreibung“, „Benennung“, „Übersetzung“, „Interview“, „Beratung“, „Diskussion“, „Begrüßung“ etc. Mit diesen Beispielwerten wollen wir keine Vorentscheidung über eine im Einzelfall anzuwendende Systematik treffen.

Im nächsten Schritt wird die Zahl der Sprecher notiert, wobei verbal beteiligte Forscher / Aufnahmeleiter mitgezählt werden sollten, was man dann im Feld „Forscherbeteiligung“ verdeutlichen kann. Für „Forscherbeteiligung“ haben wir die Werte „Verbal beteiligt“, „Nicht verbal beteiligt“ und „Nicht vorhanden“ (für „Forscher nicht anwesend“) vorgesehen. Bei Ereignissen, an denen mehrere Forscher teilgenommen haben, kann es nötig werden, die beteiligten Forscher zu benennen. In solchen Fällen können die Namen den Werten „Verbal beteiligt“ bzw. „Nicht verbal beteiligt“ vorangestellt werden.

„Elizitierung“ ist eine Technik zur Erhebung sprachlicher Daten, bei der die Informanten systematisch zu Äußerungen veranlasst werden. Wir haben die Werte „Elizitiert“ und „Nicht elizitiert“ vorgesehen.

„Mediale_Realisierung“ steht für den jeweiligen Kommunikationskanal (wie z.B. „Face to Face“, „Telefon“, „Hörfunk“). Für das Feld „Öffentlichkeitsgrad“ werden die Werte „Öffentlich“ und „Nicht öffentlich“ bereitgestellt.

Über Instruktionen und ggf. auch über Materialien, die den Sprechern zur Lösung bestimmter Aufgaben vorgelegt wurden, kann man im Feld „Vorgaben“ informieren.

Die Position eines Sprechereignisses im Ereignis kann relevant sein, wenn Segmente des aufgezeichneten kommunikativen Gehalts eines Ereignisses betrachtet werden. In solchen Fällen kann man hier die Zusammenhänge beschreiben. Eine mögliche Positionsbeschreibung wäre: „Beginnt unmittelbar nach der Begrüßung der Beteiligten und endet vor der ersten längeren Pause“. Im Feld „Sprachen“ sind die im Sprechereignis verwendeten Sprachen zu verzeichnen.

5.5.2. Inhalt

Im Komplex „Inhalt“ ist eine Beschreibung des Sprechereignisses vorgesehen. Wir nennen im Folgenden zwei fiktive Beispiele: „Großvater erzählt seinem Enkel ein altes türkisches Märchen mit dem Titel xyz, das er in seiner Jugend in seinem türkischen Heimatdorf gehört hat.“, „Anwalt berät einen Klienten über rechtliche Möglichkeiten im Konflikt mit dessen Nachbarn.“

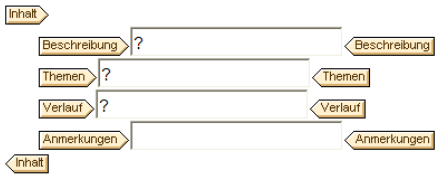

Abb. 26, Sprechereignis - Inhalt

Themenangaben sollten stichwortartig sein (z.B. „Kochrezepte“, „Lebenslauf, „Kindheitserinnerungen“, „sprachliche Entwicklungen“). Informationen über den Verlauf des Sprechereignisses kann man im gleichnamigen Feld notieren. Das können einfache Hinweise wie z.B. „Sehr turbulent“ oder komplexere Angaben über die Entwicklung sein.

5.5.3. Sprecher

Im Rahmen des Ereignisschemas werden hauptsächlich Sprecherdaten erfasst, die für das dokumentierte Sprechereignis spezifisch sind. Für allgemeine Informationen über Sprecher steht ein separates (XML-) Schema zur Verfügung (vgl. Abschnitt 6.). Der Sprecherkomplex ist fakultativ, für den Fall, dass keine Sprecherdaten erhoben wurden. Da an einem Sprechereignis mehrere Sprecher beteiligt sein können, wurde der Komplex als iterativ gekennzeichnet.

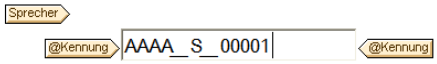

Abb. 27, Sprecher - Kennung

Die an erster Stelle der Sprecherdaten erwartete Kennung enthält die vierstellige Korpuskennung, den Kennbuchstaben S (für „Sprecher“) und eine fünfstellige laufende Nummer: AAAA_S_00001, AAAA_S_00002 usw. Da mittels dieser Kennung eine Verbindung zu den allgemeinen Sprecherdaten (vgl. Abschnitt 6.) hergestellt wird, muss sie mit der dort für den Sprecher eingetragenen Kennung übereinstimmen.

5.5.3.1. Basisdaten

In einem Fall weichen wir von unserem Prinzip, vom Sprechereignis unabhängige Sprecherdaten in einem separaten Schema zu speichern, ab. Wir übernahmen „Geschlecht“ in den Sprecherblock des Ereignisschemas, um Nutzern, die an dieser elementaren Angabe interessiert sind, ein Umschalten auf die Dokumentation allgemeiner Sprecherdaten zu ersparen.

Das Element „Rolle“ ist für Angaben über die Beteiligungsrolle des jeweiligen Sprechers in dem zu dokumentierenden Sprechereignis vorgesehen. In den Basisdaten gibt es die Möglichkeit, Informationen über Besonderheiten eines Sprechers im Sprechereignis, wie z.B. „War anfangs sehr nervös“ oder „Stellte sich in hohem Maße auf seinen Gesprächspartner ein“, „Vermied Blickkontakte“ etc., zu erfassen. Ein Feld für sprachliche Besonderheiten ist in den „Sprachdaten“ (vgl. 5.5.3.2.) enthalten.

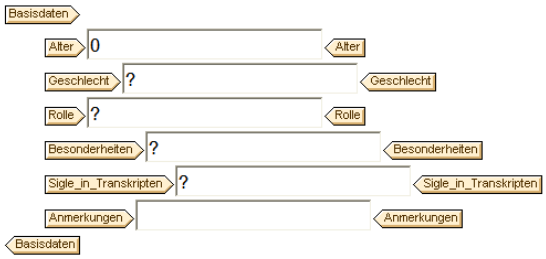

Abb. 28, Sprecher - Basisdaten

Am Ende der Basisdaten soll die in Transkripten verwendete Sprechersigle notiert werden. An dieser Stelle möchten wir darauf hinweisen, dass für einen Sprecher nicht mehrere Siglen vergeben werden sollten.

5.5.3.2. Sprachdaten

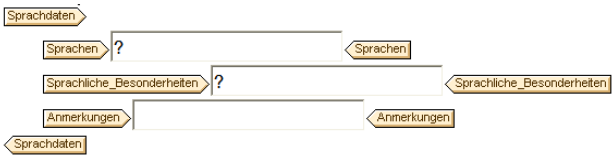

Abb. 29, Sprecher - Sprachdaten

„Sprachdaten“ ist ein Bereich zur Erfassung von Angaben über die von einem Sprecher im jeweiligen Sprechereignis gesprochene(n) Sprache(n) und seine sprachlichen Besonderheiten.

Zunächst sind die Namen der verwendeten Sprachen (z.B. „Deutsch ; Englisch ; Türkisch“) einzutragen. Unter dem Stichwort „Sprachliche Besonderheiten“ können weitere für das Sprechereignis charakteristische sprachliche Merkmale notiert werden. Die folgenden Beispiele stammen aus der Dokumentation des Korpus „Emigrantendeutsch in Israel“ von Anne Betten: „Wiener Verkehrsmundart mit einzelnen österreichischen Dialekteigenheiten. Lebhafter Erzählstil.“(IS010) „Sehr gewandtes, nuancenreiches Sprechen; überwiegend standardsprachlich korrekt formulierend, aber themenabhängig variierend von sehr verhaltenen bis zu drastisch-lebendigen Ausdrucksweisen. Stimme stark modulierend.“ (IS015) [13]

5.6. Sprechereignisspezifische Aufnahme

Wir nehmen an, dass es zu jedem dokumentierten Sprechereignis (mindestens) eine Aufnahme (SE-Aufnahme) gibt, die in einem bestimmten Verhältnis zu einer Quellaufnahme steht. Meist haben wir es mit einem spezifischen Teil in einer Quellaufnahme bzw. einer Kopie dieses spezifischen Teils zu tun. Es kommt allerdings auch vor, dass SE-Aufnahmen mit Quellaufnahmen vollkommen übereinstimmen. In solchen Fällen handelt es sich bei der angenommenen SE-Aufnahme um ein Konstrukt, das es uns erlaubt, an dieser Stelle einen Bezug zu einer Quellaufnahme herzustellen. Da mehrere SE-Aufnahmen vorliegen können, haben wir diesen Komplex im Schema als iterativ gekennzeichnet.

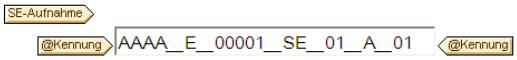

Abb. 30, SE-Aufnahme - Kennung

Die Kennung für eine SE-Aufnahme umfasst die Kennung des Sprechereignisses, den Kennbuchstaben A (für Aufnahme) und eine zweistellige laufende Nummer. Ein Beispiel dafür finden Sie in Abb. 30.

5.6.1. Basisdaten

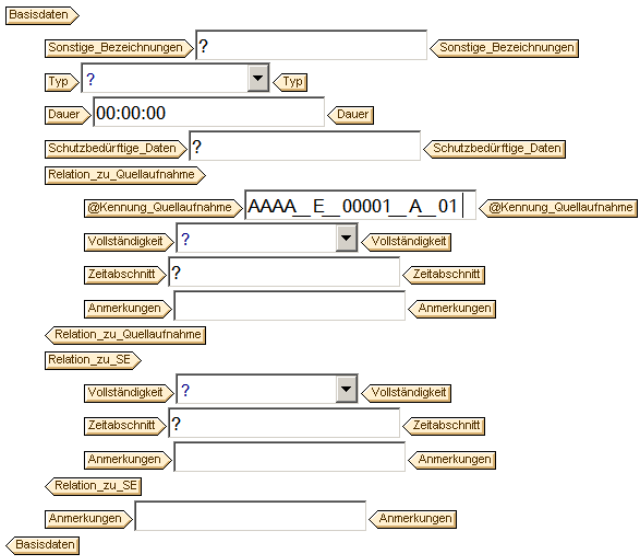

Abb. 31, SE-Aufnahme - Basisdaten (1)

An erster Stelle der Basisdaten steht das Feld „Sonstige_Bezeichnungen“. Damit sind im Projekt vergebene Kurzbezeichnungen gemeint. SE-Aufnahmen können unterschiedlichen Typs sein: Audioaufnahme, Videoaufnahme und ggf. auch Tonkopie von Videoaufnahme.

Für Angaben über die Dauer der jeweiligen Aufnahme wurde ein fakultatives Zeitfeld mit dem Format hh:mm:ss vorbereitet. Aufnahmen können Daten enthalten, die nach dem Willen der Urheber und aus datenschutzrechtlichen Gründen Außenstehenden nicht kenntlich werden dürfen, wie z.B. persönliche Sprecherdaten. Für entsprechende Informationen wurde das fakultative Feld „Schutzbedürftige_Daten“ bereitgestellt. Wenn die dokumentierten SE-Aufnahmen mit den Quellaufnahmen vollständig übereinstimmen, können diese beiden Felder bei der Erstellung projektspezifischer Schemata übergangen werden.

Unter der Überschrift „Relation_zu_Quellaufnahme“ kann man Informationen über das Verhältnis der jeweiligen SE-Aufnahme zu einer oder mehreren Quellaufnahmen erfassen. Dieses Modul ist iterativ. An erster Stelle ist die Kennung der Quellaufnahme (z.B. AAAA_E_00001_A_01) einzutragen. Eine SE-Aufnahme kann mit einer Quellaufnahme vollständig übereinstimmen oder ein Segment in einer Quellaufnahme sein. Diese Angaben sind im Feld „Vollständigkeit“ mit den Werten „Vollständig“ bzw. „Segment“ zu notieren. Die Angabe „Segment“ sollte im Feld „Zeitabschnitt“ präzisiert werden. Bei vollständigen Aufnahmen wird hier der Wert „Vollständig“ erwartet.

Unter der Überschrift „Relation_zu_SE“ werden Informationen über das Verhältnis der jeweiligen SE-Aufnahme zum Sprechereignis erfasst. Im Feld „Vollständigkeit“ wird eingetragen, ob eine vollständige oder eine unvollständige Aufnahme eines Sprechereignisses dokumentiert wird. Die Angabe „Unvollständig“ sollte man im Feld „Zeitabschnitt“ präzisieren. „Zeitabschnitt“ meint den in der jeweiligen SE-Aufnahme aufgezeichneten Zeitabschnitt des Sprechereignisses. Wenn die genaue Zeit nicht zu ermitteln ist, kann hier darüber informiert werden, welcher Abschnitt des Sprechereignisses in der jeweiligen Aufnahme festgehalten ist, z.B. „1. Abschnitt“, „2. Abschnitt“ usw. Bei vollständigen Aufnahmen wird der Wert „Vollständig“ erwartet.

5.6.2. Technische Fassungen

SE-spezifische Kopien von Quellaufnahmen liegen in ganz bestimmten technischen Fassungen vor. Die Komponenten „Analoge_Fassung“ und „Digitale_Fassung“ sind fakultativ und iterativ. Sollten alle SE-Aufnahmen eines Korpus mit den Quellaufnahmen identisch oder keine SE-spezifischen Kopien vorhanden ein, können beide Komponenten bei der Erstellung korpus-spezifischer Schemata übergangen werden.

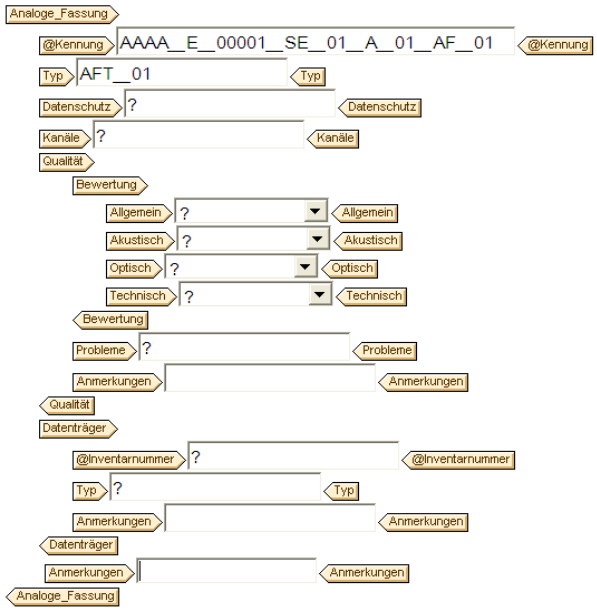

Abb. 32, SE-Aufnahme - Analoge Fassung

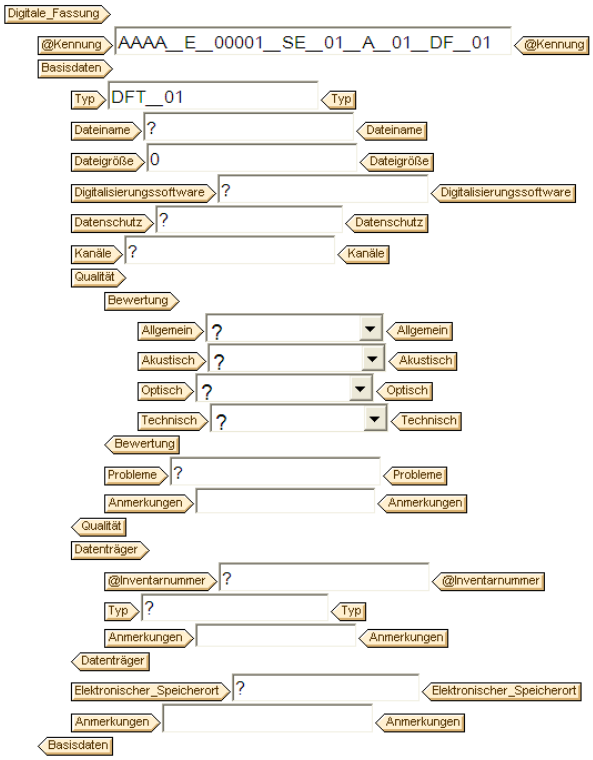

Abb. 33, SE-Aufnahme - Digitale Fassung – Kennung und Basisdaten

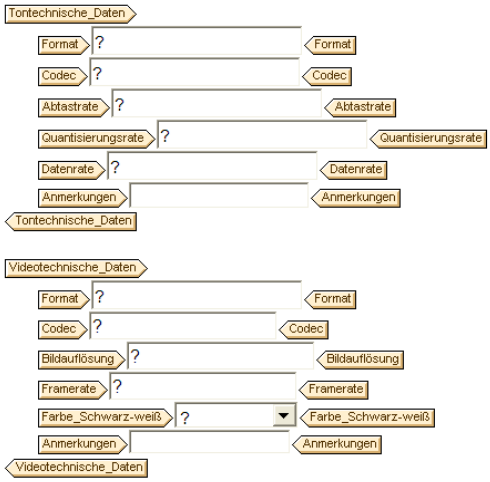

Abb. 34, SE-Aufnahme - Digitale Fassung – Ton- und Videotechnische Daten

Die Strukturen dieser Komponenten stimmen mit den entsprechenden Strukturen im Komplex Quellaufnahmen überein. Erläuterungen dazu finden Sie in Abschnitt 5.3.3. Die Kennungen sind allerdings ebenenspezifisch und setzen sich hier zusammen aus der Kennung der SE-Aufnahme, dem Kürzel AF (für „Analoge_Fassung“) bzw. DF (für „Digitale_Fassung“) und einer zweistelligen Nummer (vgl. Abb. 32 und 33).

5.6.3. Archivierung und Distribution

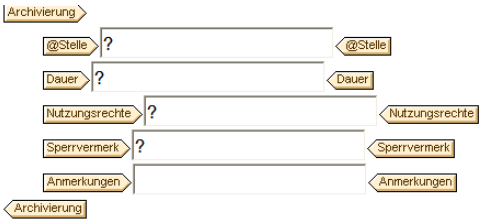

Abb. 35, SE-Aufnahme - Archivierung

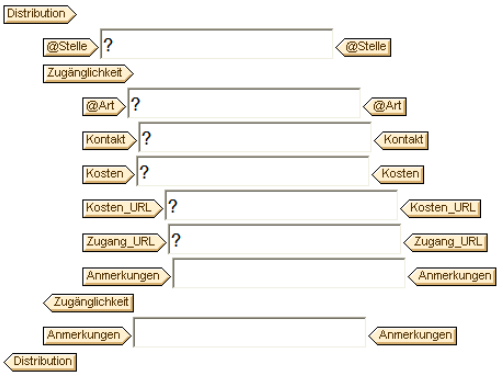

Abb. 36, SE-Aufnahme - Distribution

Auch mit den Modulen „Archivierung“ und „Distribution“ haben wir Sie schon bekannt gemacht. Sie werden in der Beschreibung aller Korpusbestandteile verwendet und in Abschnitt 5.3.4. eingeführt. Für Fälle, in denen alle SE-Aufnahmen eines Korpus mit den Quellaufnahmen identisch sind, wurden diese Verwaltungsdaten für SE-Aufnahmen als fakultativ gekennzeichnet.

5.7. Transkript

Der gesamte Transkriptkomplex ist fakultativ. Da zu einer SE-Aufnahme mehrere Transkripte vorliegen können, haben wir ihn im Schema als iterativ gekennzeichnet.

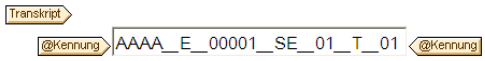

Abb. 37, Transkript - Kennung

An erster Stelle der Transkriptdaten wird eine Kennung vergeben, die die jeweilige Sprecher-eigniskennung, das Kürzel T (für „Transkript“) und eine zweistellige laufende Nummer umfasst. Ein Beispiel finden Sie in Abb. 37.

5.7.1. Basisdaten

Die Felder „Sonstige_Bezeichnungen“ und „Titel“ in den Transkript-Basisdaten stehen für im Projekt vergebene Kurzbezeichnungen und Transkripttitel.

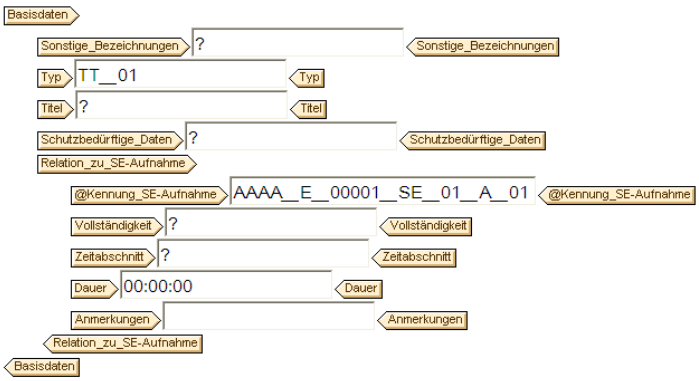

Abb. 38, Transkript - Basisdaten

Transkripte können typisiert werden, wobei die Extension sowie Art und Anzahl der Annotationen relevant werden können. Für die Bezeichnung von Transkripttypen wird das Kürzel TT in Verbindung mit einer zweistelligen Nummer verwendet.

Transkripte können Daten enthalten, die nach dem Willen der Urheber und aus datenschutzrechtlichen Gründen Außenstehenden nicht kenntlich werden dürfen, wie z.B. persönliche Sprecherdaten, über die man im Feld „Schutzbedürftige_Daten“ informieren kann.

Im Abschnitt „Relation_zu_SE-Aufnahme“ wird zunächst die Kennung der transkribierten Aufnahme abgefragt. Im Feld „Vollständigkeit“ soll notiert werden, ob es sich um ein vollständiges oder ein Teiltranskript handelt. „Zeitabschnitt“ meint den Zeitabschnitt des transkribierten Teils der SE-Aufnahme. Wenn sich ein Transkript auf unterschiedliche Teile einer SE-Aufnahme bezieht (auf Auslassungen wird dann i.d.R. im Transkripttext hingewiesen), sind hier mehrere Werte zu erwarten. Für Angaben über die Dauer einer transkribierten Aufnahme bzw. eines Aufnahmeausschnittes wurde ein Zeitfeld mit dem Format hh:mm:ss vorbereitet. Bei vollständig transkribierten Aufnahmen sollte die Angabe zur Dauer der SE-Aufnahme übernommen werden.

5.7.2. Annotation

Wir verwenden die Bezeichnung „Annotation“ für inhaltlich und formal charakterisierte Ebenen eines Transkripts, wie z.B. Aufzeichnungen des Wortlauts in orthographischer, literarischer oder phonetischer Umschrift, syntaktische Angaben, Notationen suprasegmentaler oder nonverbaler Phänomene, Übersetzung des Wortlautes etc. [14] Da u.U. mehrere Annotationen pro Transkript zu dokumentieren sind, wurde der Komplex im Schema als iterativ gekennzeichnet. [15]

Abb. 39, Annotation - Typ

Zur einfachen Benennung unterschiedlicher Annotationen haben wir eine Typenbezeichnung eingeführt. Diese Bezeichnung besteht aus dem Kürzel „ANT“ (für „Annotation_Typ“) und einer zweistelligen laufenden Nummer. Ein Beispiel finden Sie in Abb. 39.

5.7.2.1. Basisdaten

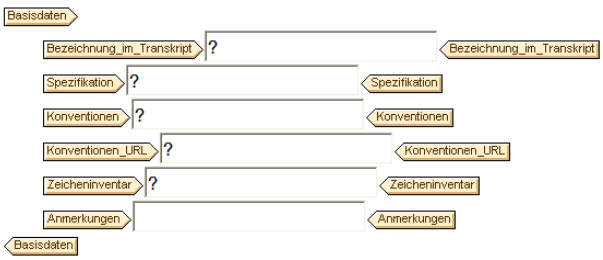

Abb. 40, Annotation - Basisdaten

An erster Stelle der Basisdaten kann eine im Transkript enthaltene Bezeichnung für die jeweilige Annotation eingetragen werden. Im Feld „Spezifikation“ wird eine Charakterisierung der Annotation erwartet. Hier sollte man über den Gegenstand (z.B. „Wortlaut“), die Umschrift (z.B. „Literarisch“) und die Reichweite (z.B. „Ohne Interviewerbeiträge“, „Nur für Sprecher XY“) informieren.

Auf die Konventionen für die jeweilige Annotation kann man im gleichnamigen Feld hinweisen. Beispiele für solche Hinweise wären: „Projektspezifisch“, „DIDA, Version vom Januar 2001“, „GAT“ etc. Über eine URL kann man ggf. direkt auf diese Konventionen zugreifen. Unter „Zeicheninventar“ ist das Inventar an Schriftzeichen zu verstehen, das bei der Wiedergabe des Wortlautes verwendet wurde. Das sind i.d.R. standardisierte Inventare wie z.B. der IPA-Zeichensatz oder ein spezifisches Alphabet.

5.7.2.2. Erstellung

Die Erstellung einer Annotation umfasst nach unserem Verständnis neben der Ersterstellung auch Ergänzungen und Korrekturen. Da man eine Annotation mehrmals überarbeiten kann, wurde die Komponente „Erstellung“ im Schema als iterativ gekennzeichnet.

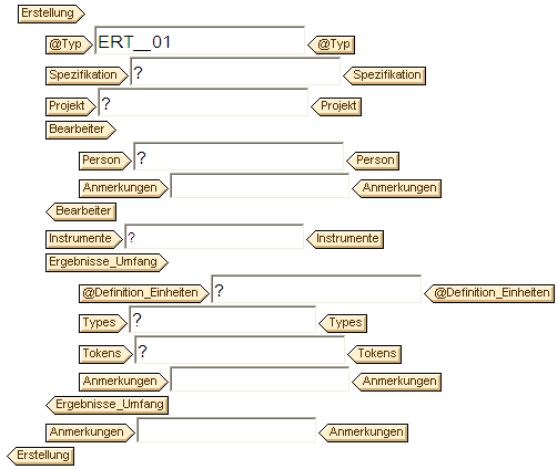

Abb. 41, Transkript - Annotation - Erstellung

Um verschiedene Erstellungsprozesse einfach benennen zu können, wurde auch an dieser Stelle eine Typenbezeichnung eingefügt. Sie setzt sich zusammen aus dem Kürzel ERT (für „Erstellung_Typ“) und einer zweistelligen Nummer.

Das Feld „Spezifikation“ wurde für Informationen über die Art der Erstellung (z.B. „Ersterfassung“, „1. Korrektur“, „Endkorrektur“, „Überarbeitung für Publikation xy“) und mögliche besondere Umstände (z.B. „halbautomatisch“) bereitgestellt. Die Werte dieses Feldes sind grundlegend für die Typisierung. Für den Namen eines Erstellungsprojekts wurde das Feld „Projekt“ eingefügt. Im Feld „Instrumente“ kann man Angaben über den genutzten Editor und evtl. weitere Hinweise auf die Systemumgebung notieren.

Informationen über den Umfang der Ergebnisse einer Erstellung sollten eine Definition der gezählten Einheiten, Angaben über die Anzahl unterschiedlicher Einheiten (Types) und die Anzahl aller gezählten Einheiten (Tokens) umfassen. Der Komplex wurde im Schema als iterativ gekennzeichnet.

5.7.2.3. Alignment

Wir verwenden die Bezeichnung „Alignment“ in der Dokumentation für die Text-Ton-Synchronisation, also die Koppelung von Aufnahmen und Transkripten auf Phon-, Phonem-, Wort- oder Phrasenbasis, wobei Transkriptsegmenten Zeitmarken zugeordnet werden. Die Komponente ist fakultativ und iterativ.

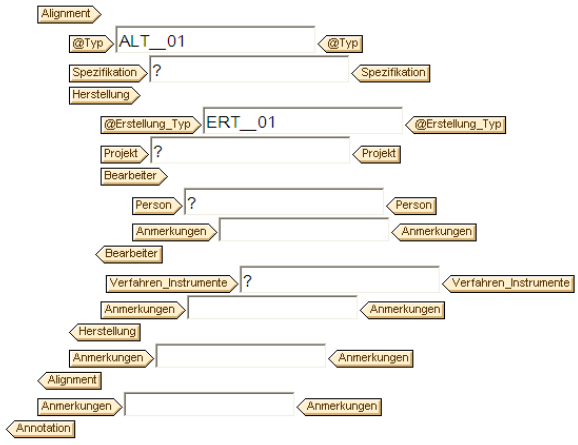

Abb. 42, Transkript - Annotation - Alignment

Alignmentprozessen werden Typenbezeichnungen zugeordnet, die aus dem Kürzel ALT (für „Alignment_Typ“) und einer zweistelligen Nummer bestehen. Grundlage für die Typisierung sind die Angaben im Feld „Spezifikation“, wo über die alignierten Segmente (z.B. „Phonweise“, „Wortweise“) informiert wird.

Die Komponente „Herstellung“ ist iterativ. Zunächst wird der Typ der Erstellung verzeichnet, dessen Ergebnisse aligniert wurden. Im Feld „Projekt“ wird der Namen des Projekts, in dem das Alignment vorgenommen wurde, erwartet. Im Feld „Verfahren_Instrumente“ kann man Angaben darüber machen, ob manuell oder automatisch aligniert wurde, auf die genutzte Software hinweisen und ggf. weitere Informationen über die Systemumgebung erfassen.

5.7.3. Technische Fassungen

Transkripte können in verschiedenen technischen Fassungen vorliegen. Um diese dokumentieren zu können, haben wir auch an dieser Stelle die Komponenten „Analoge_Fassung“ und „Digitale_Fassung“ eingerichtet. Beide Komponenten sind fakultativ und iterativ, können aber nicht beide bei der Erstellung eines korpusspezifischen Schemas übergangen werden.

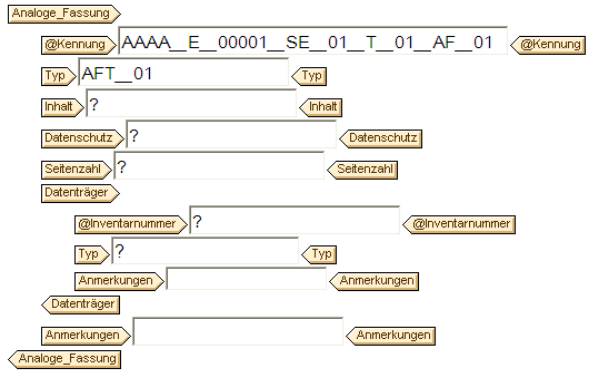

Abb. 43, Transkript - Analoge Fassung

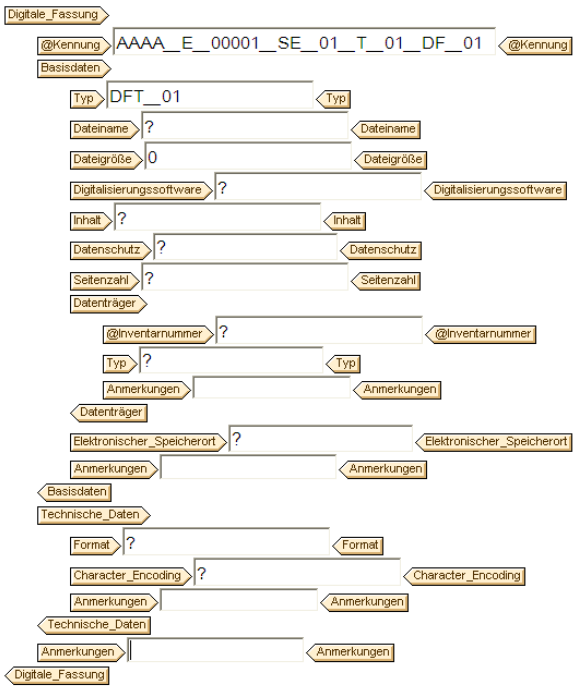

Abb. 44, Transkript - Digitale Fassung

Es folgt ein Feld für Informationen über den jeweiligen Typ der technischen Fassungen. Für die Typisierung analoger Fassungen sind die Werte für Datenschutz und Inhalt relevant, bei der Typisierung digitaler Fassungen auch technischen Daten (wie z.B. das Dateiformat).

Im iterativen Feld „Inhalt“ sollte man die Annotationstypen sowie die Typen der Erstellungs- und Alignmentprozesse notieren, deren Ergebnisse in der jeweiligen Fassung gespeichert sind. Im Anschluss daran wird eine Information über Maßnahmen zum Datenschutz (wie z.B. die Maskierung von Personennamen) erwartet. Für den Fall, dass seitenformatierte Transkripte vorliegen und der Seitenumfang zu dokumentieren ist, haben wir das Feld „Seitenzahl“ eingefügt.

Die anderen Komponenten dieser Module stimmen mit den entsprechenden Komponenten für Zusatzmaterial überein. Erläuterungen dazu finden Sie in Abschnitt 5.4.2.

5.7.4. Archivierung und Distribution

Die Module „Archivierung“ und „Distribution“ sind in der Dokumentationsstruktur für alle Korpusbestandteile enthalten ist. Die Erläuterungen in Abschnitt 5.3.4. gelten auch für Transkripte.

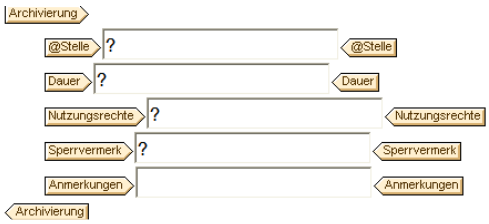

Abb. 45, Transkript - Archivierung

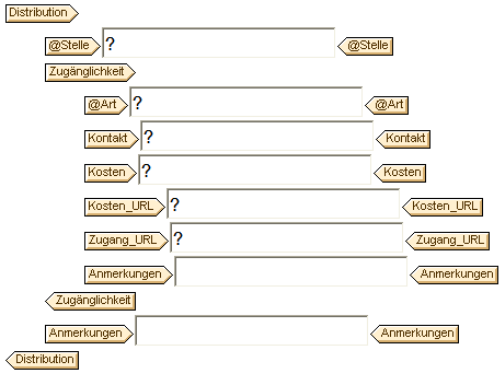

Abb. 46, Transkript - Distribution

5.8. Zusatzmaterial

Unter „Zusatzmaterial“ auf der Sprechereignisebene verstehen wir solche Dokumente, die zusätzlich zu Aufnahmen und Transkripten vorhanden sein können. Für ein Sprechereignis spezifische Unterlagen, wie z.B. schriftliche Vorbereitungen eines Interviews oder längere Aufzeichnungen über den Verlauf, sollten als Zusatzmaterial dokumentiert werden. Der gesamte Komplex „Zusatzmaterial“ ist fakultativ. Da zu einem Sprechereignis mehrere Dokumente vorliegen können, wurde er im Schema als iterativ gekennzeichnet.

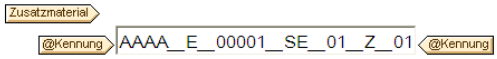

Abb. 47, Sprechereignis - Zusatzmaterial –Kennung

Die hier geforderte Kennung setzt sich zusammen aus der Kennung des Sprechereignisses, dem Kennbuchstaben Z (für „Zusatzmaterial“) und einer zweistelligen laufenden Nummer. Ein Beispiel finden Sie in Abb. 47.

5.8.1. Basisdaten

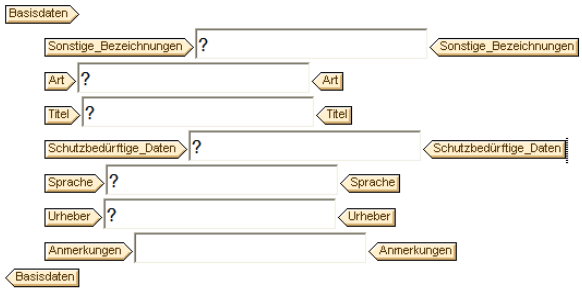

Abb. 48, Sprechereignis - Zusatzmaterial - Basisdaten

Das Modul „Basisdaten“ ist für Zusatzmaterial auf der Ereignisebene und der Sprechereignisebene gleich strukturiert. Eine Beschreibung der Struktur finden Sie in Abschnitt 5.4.1.

5.8.2. Technische Fassungen

Auch die Strukturen der Dokumentation technischer Fassungen von Zusatzmaterial auf Ereignis- und auf Sprechereignisebene stimmen überein. In beiden Fällen haben wir die Komponenten „Analoge_Fassung“ und „Digitale_Fassung“ im Schema als fakultativ und iterativ gekennzeichnet. Bei der Erstellung korpusspezifischer Schemata muss wenigstens eine Komponente gewählt werden.

Die Kennungen der technischen Fassungen sind allerdings ebenenspezifisch. An dieser Stelle besteht die Kennung aus der Kennung des Zusatzmaterials auf Sprechereignisebene, dem Kürzel AF (für „Analoge_Fassung“) bzw. DF (für „Digitale_Fassung“) und einer zweistelligen Nummer. Beispiele sind in den Abb. 49 und 50 enthalten. Erläuterungen der anderen Felder in diesen Modulen finden Sie in Abschnitt 5.4.2.

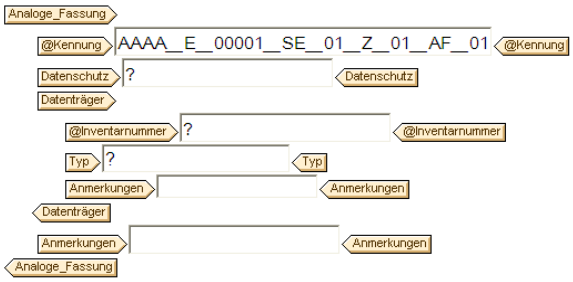

Abb. 49, Sprechereignis - Zusatzmaterial - Analoge Fassung

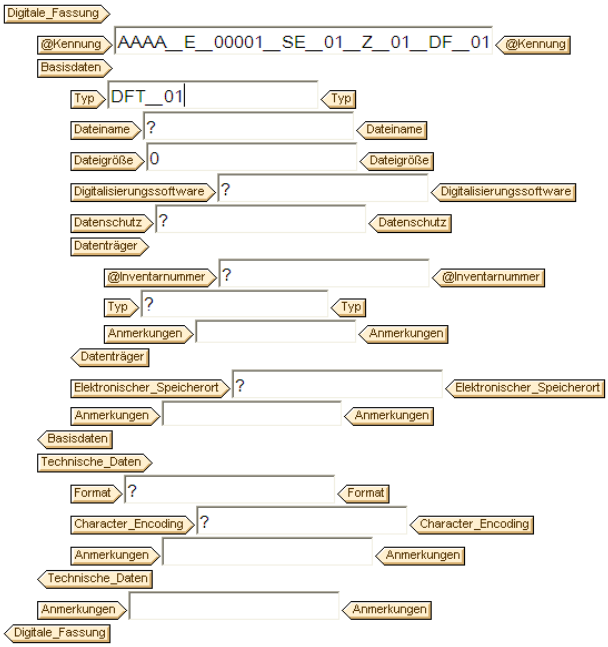

Abb. 50, Sprechereignis - Zusatzmaterial - Digitale Fassung

5.8.3. Archivierung und Distribution

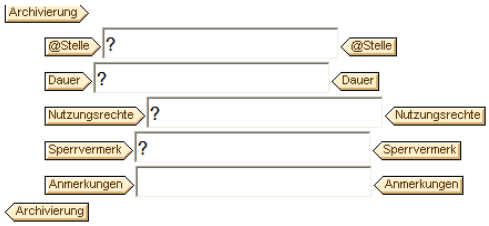

Abb. 51, Sprechereignis - Zusatzmaterial - Archivierung

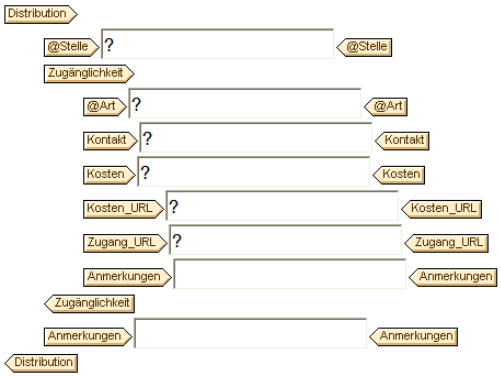

Abb. 52, Sprechereignis - Zusatzmaterial - Distribution

„Archivierung“ und „Distribution“ sind Bausteine, die in der Dokumentation aller Korpusbestandteile gebraucht werden. Die Erläuterungen in Abschnitt 5.3.4. gelten auch für Zusatzmaterial auf Sprechereignisebene.

5.9. Dokumentationsgeschichte

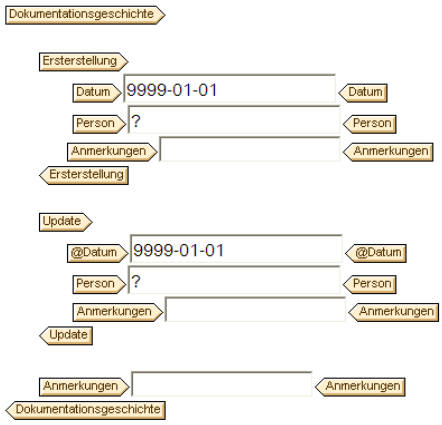

Abb. 53, Dokumentationsgeschichte

Informationen über Arbeitsstand und Bearbeiter der Dokumentation werden bei der manuellen Dateneingabe automatisch in einer (Oracle-)Datenbank gespeichert, sollten nach unserer Vorstellung jedoch auch in den Dokumenten sichtbar sein..

6. Generisches Schema für die Dokumentation allgemeiner Sprecherdaten

Neben einem generischen Schema für die Dokumentation von Korpusbestandteilen auf der Ereignisebene (vgl. Abschnitt 5.) gibt es ein Schema für die Dokumentation ereignis- und sprechereignisübergreifender Sprecherdaten. Auch für dessen Darstellung haben wir ein projektneutrales Erfassungsformular erstellt, an dem wir uns im Folgenden orientieren.

Die Kategorie „Sprecher“ dient als Startknoten eines (XML-)Schemas, das folgende Informationen vorsieht: Angaben über den jeweiligen Sprecher (Basisdaten, Ortsdaten, Sprachdaten), Beziehungen dieses Sprechers zu anderen Sprechern, sonstige Bezugspersonen des Sprechers, Vereinbarungen über Datenschutz und Nutzungsrechte, Zusatzmaterial auf Sprecherebene sowie eine Dokumentationsgeschichte.

Auch dieses Schema enthält obligatorische und fakultative Komponenten. Obligatorische Komponenten sind in allen projektspezifischen Subschemata zu berücksichtigen, fakultative Komponenten stehen zur Wahl und müssen in den Subschemata nicht verwendet werden. Wenn man sie verwendet, sind alle Kennungsfelder und die mit ? gekennzeichneten Felder zu bearbeiten.

Eingaben für fehlende Daten in diesen Feldern sind standardisiert: „Nicht dokumentiert“ bedeutet: Es kann ein Datum geben, das bei der Datenerfassung nicht bekannt ist. Ein Beispiel dafür wäre: „Ethnische Zugehörigkeit: Nicht dokumentiert“ - zu lesen als: „Die ethnische Zugehörigkeit des Sprechers ist nicht bekannt.“ „Nicht vorhanden“ bedeutet: Es gibt kein Datum. Ein Beispiel dafür wäre: „Aktuell ausgeübter Beruf: Nicht vorhanden“ - zu lesen als: „Der Sprecher ist nicht berufstätig.“

Das an vielen Stellen vorgesehene Feld „Anmerkungen“ ist für Anmerkungen zu Angaben in anderen Feldern und für nicht kategorisierte Angaben vorgesehen. Das Feld kann leer bleiben.

Einzelne Komponenten des Schemas wurden als iterativ gekennzeichnet, d.h. dass sie bei der korpusspezifischen Datenerfassung vervielfältigt werden können.

Die nachfolgenden Abbildungen stammen aus einem projektneutralen Erfassungsformular, das zu Demonstrationszwecken angelegt wurde.

6.1. Sprecher

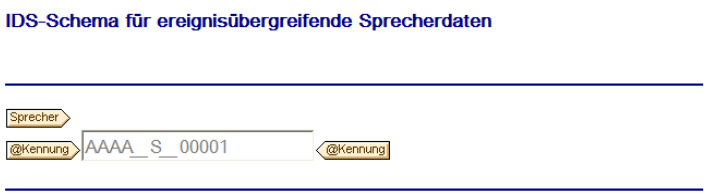

Abb. 54, Sprecher - Kennung

An erster Stelle der Sprecherdaten steht eine Sprecher-Kennung, die eine vierstellige Korpuskennung, den Kennbuchstaben S (für „Sprecher“) und eine fünfstellige laufende Nummer umfasst. Ein Beispiel finden Sie in Abb. 54. Diese Kennung ist auch im Sprecherblock des Ereignisschemas (vgl. 5.5.3.) zu verzeichnen.

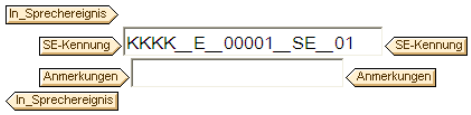

Abb. 55, Sprecher – in Sprechereignisereignis

Das iterative Feld „SE-Kennung“ soll die Kennung des Sprechereignisses aufnehmen, in dem der hier dokumentierte Sprecher aktiv war. Mit Hilfe eines Links, der jeder Kennung hinterlegt wird, wird eine Verbindung zur Dokumentation des Sprechereignisses im entsprechenden Ereignisdokument hergestellt.

6.1.1. Basisdaten

An erster Stelle der Sprecher-Basisdaten steht das Feld „Sonstige_Bezeichnungen“. Damit sind projektinterne Kurzbezeichnungen des Sprechers gemeint. Die Felder „Name“ und „Früherer_Name“ wurden mit dem Wert „Anonym“ vorbelegt und werden nur dann anders genutzt, wenn Sprechernamen Außenstehenden kenntlich werden dürfen. Für den Fall, dass maskiert wurde, haben wir das Feld „Pseudonym“ eingerichtet.

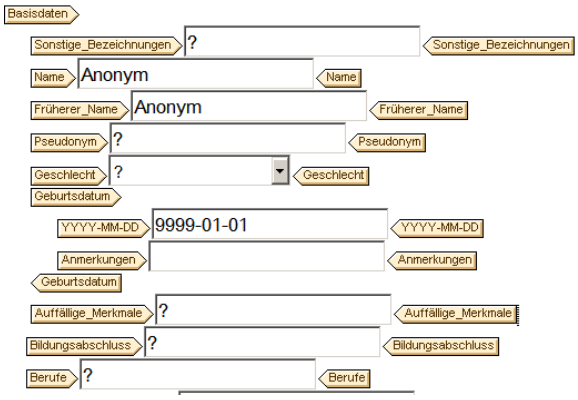

Abb. 56, Sprecher - Basisdaten (1)

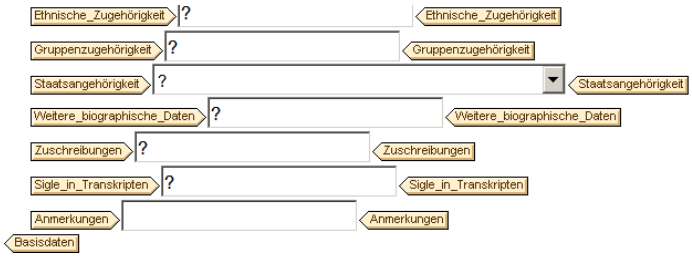

Abb. 57, Sprecher - Basisdaten (2)

Im Anschluss daran wird das Geschlecht erfasst. Es folgt das Geburtsdatum, wobei das zugehörige Feld „Anmerkungen“ die Möglichkeit bietet, auf ungenaue Angaben hinzuweisen. Unter der Bezeichnung „Auffällige Merkmale“ kann man z.B. über körperliche Behinderungen oder auffällige Kleidung informieren. Es folgen Angaben über Bildungsabschluss und Berufe.

In das Feld „Ethnische_Zugehörigkeit“ sollten Selbsteinschätzungen der Sprecher eingetragen werden. Im Feld „Gruppenzugehörigkeit“ kann man Informationen über die Zugehörigkeit eines Sprechers zu sozialen Gruppen und seine Positionen in diesen Gruppen, wie z.B. „Mitglied des Gemeinderats“ oder „Vorsitzende des örtlichen Tierschutzvereins“, ablegen. Für das iterative Feld „Staatsangehörigkeit“ gibt es eine ISO-Länderliste.

Weitere biographische Daten, für die keine Kategorien bereitgestellt wurden, können im gleichnamigen Feld notiert werden. Mit „Zuschreibungen“ sind Attribute gemeint, die sich der Sprecher selbst zuschreibt und / oder die dem Sprecher von anderen zugeschrieben werden, wie z.B. „Manager des Jahres“ oder „Bürgerschreck“.

Das Feld „Sigle_in_Transkripten“ muss vermutlich nicht erläutert werden. Da in älteren Korpora z.T. mehrere Siglen pro Sprecher verwendet wurden, haben wir das Feld auch im Sprecherblock des Ereignisschemas (vgl. 5.5.3.) bereitgestellt. Wir möchten an dieser Stelle noch einmal darauf hinweisen, dass für einen Sprecher nicht mehrere Siglen vergeben werden sollten.

6.1.2. Sprecher - Ortsdaten

Der Komplex „Ortsdaten“ ist fakultativ. Damit Informationen über mehrere sprachlich relevante Orte erfasst werden können, wurde er im Schema als iterativ gekennzeichnet.

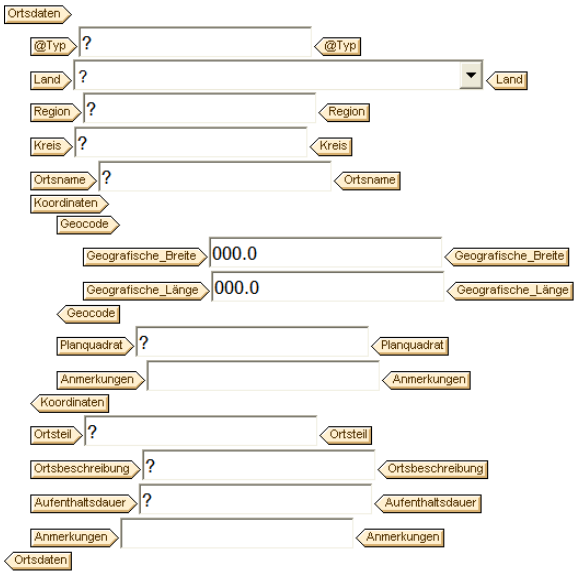

Abb. 58, Sprecher - Ortsdaten

An erster Stelle der Ortsdaten wird der Ortstyp abgefragt. Für dieses Feld sind Werte wie „Geburtsort“, „Wohnort“, „Arbeitsort“ relevant. Für das Feld „Land“ gibt es eine ISO-Liste. Das Feld „Region“ ist für die amtliche Bezeichnung eines Bundeslandes, eines Kantons oder einer Provinz vorgesehen. In die Felder „Kreis“ und „Ortsname“ sollen ebenfalls amtliche Bezeichnungen eingetragen werden.

Die Komponente „Koordinaten“ ist fakultativ. Wenn sie von einem Projekt gewählt wird, ist entweder der Geocode oder das Planquadrat (Kategorie im DSAv-Katalog [9]) des Ortes zu verzeichnen. Der „Geocode“ umfasst die Felder „Geographische_Breite“ und „Geographische_ Länge“. Die Felder „Ortsteil“ und „Aufenthaltsdauer“ müssen vermutlich nicht erläutert werden. Weitere Informationen über den sprachlich relevanten Ort kann man im Feld „Ortsbeschreibung“ erfassen.

6.1.3. Sprecher - Sprachdaten

Der fakultative Bereich „Sprachdaten“ umfasst die drei fakultativen Komplexe „Sprachkenntnisse“, „Sprachproduktion“ und „Sprachgebrauch“.

6.1.3.1. Sprachkenntnisse

Um mehrsprachigen Personen gerecht werden zu können, wurde der Komplex „Sprachkenntnisse“ als iterativ gekennzeichnet.

Nach dem Sprachnamen (z.B. „Deutsch“, „Englisch“, „Türkisch“) ist der Sprachstatus anzugeben (z.B. „Muttersprache“, „Erstsprache“, „Zweitsprache“, „1. Fremdsprache“). Die Komponente „Kenntnisgrade“ ist fakultativ, alle darin enthaltenen Felder außer den Anmerkungen werden verbindlich, wenn sie gewählt wird. In Abb. 60 sehen Sie die für die Einschätzung der Kenntnisgrade vorgegebenen Werte.

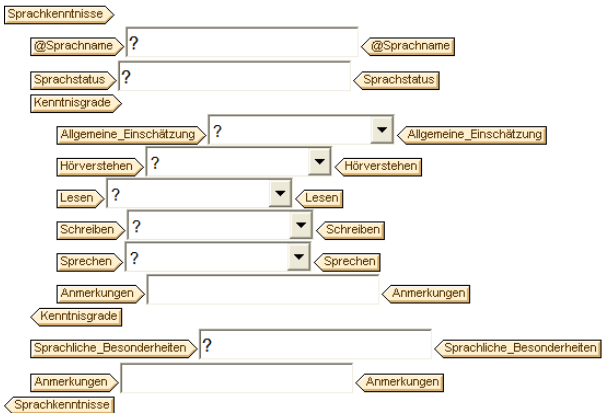

Abb. 59, Sprecher - Sprachdaten - Sprachkenntnisse

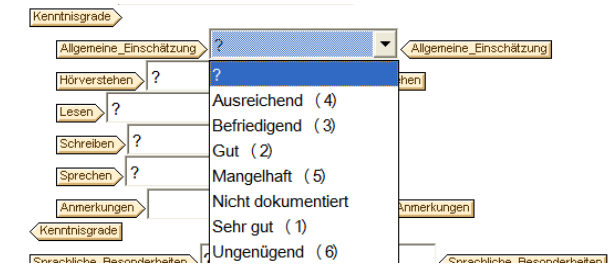

Abb. 60, Werte für die Einschätzung von Kenntnisgraden

Im Anschluss an die Kenntnisgrade kann man Informationen über sprachliche Besonderheiten, wie z.B. dialektale Aspekte, erfassen.

6.1.3.2. Sprachproduktion

Die in diesem Abschnitt vorgestellte Systematik rekurriert u.a. auf einen Sprecherfragebogen für die Erhebung „Deutsch heute“ (DH) [16]. Wir haben allerdings die in die Schemaentwicklung eingespeisten DH-Kategorien daraufhin geprüft, ob sie auch für andere Anwendungen brauchbar sind, und im Hinblick darauf überarbeitet.

Der Abschnitt Sprachproduktion ist fakultativ. Er besteht aus dem ebenfalls fakultativen Komplex „Einflussfaktoren“ und dem Element „Sprachliche_Besonderheiten“.

Der Komplex „Einflussfaktoren“ umfasst die fünf fakultativen Komponenten „Körpermaße“, „Beeinträchtigung“, „Drogen_Medikamente“, „Gebrauch_von_Hilfsmitteln“ und „Unterricht_Korrekturen“.

In Abb. 61 sehen Sie zunächst die Struktur des Moduls „Körpermaße“. Die Körpergröße wird in cm, das Körpergewicht in kg angegeben.

Das Modul „Beeinträchtigung“ wurde im Schema als iterativ gekennzeichnet. Hier können Informationen über körperliche und psychische Beeinträchtigungen erfasst werden. Wir denken dabei an Werte wie z.B. „Zahnlücke im vorderen Unterkiefer“, „Schwerhörigkeit“, „Asthma“ oder „Depression“. Im Feld „Häufigkeit_Umfang“ kann notiert werden, wie häufig bzw. in welchem Umfang eine Beeinträchtigung auftritt. Eine Information darüber, wie lange die Beeinträchtigung schon vorhanden ist, kann man im Feld „Dauer“ erfassen.

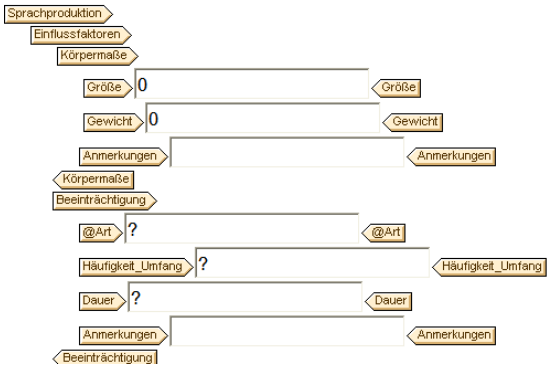

Abb. 61, Sprecher - Sprachdaten - Sprachproduktion (1)

Abb. 62 zeigt die Struktur der Komponente „Drogen_Medikamente“. Auch diese Komponente wurde im Schema als iterativ gekennzeichnet. Im Feld „Art“ kann z.B. „Nikotin“ eingetragen werden. Im Feld „Häufigkeit_Umfang“ kann man dann festhalten, welche und wie viele Rauchwaren am Tag konsumiert werden. Im Feld „Dauer“ wird vermerkt, wie lange der Sprecher schon raucht.

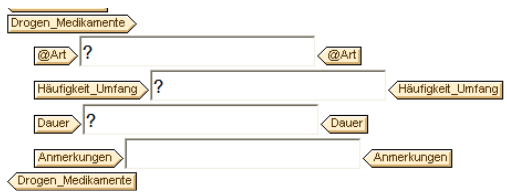

Abb. 62, Sprecher - Sprachdaten - Sprachproduktion (2)

Die Struktur der iterativen Komponente „Gebrauch_von_Hilfsmitteln“ ist in Abb. 64 dargestellt. Bei sprachlich relevanten Hilfsmitteln ist v.a. an Wörterbücher aller Art, Grammatiken, Stillehren und Kommunikationsratgeber zu denken, hier könnten aber auch andere Hilfsmittel, wie z.B. ein Hörgerät, verzeichnet werden. Für entsprechende Werte steht das Feld „Art“ bereit. Unter „Häufigkeit_Umfang“ kann vermerkt werden, wie oft ein Hilfsmittel genutzt wird. Die übrigen Felder stimmen mit den in Abb. 61 gezeigten überein.

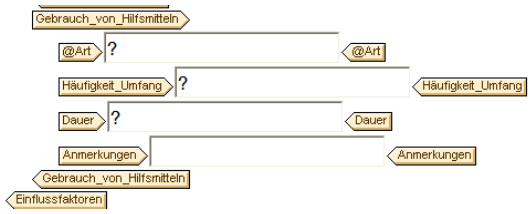

Abb. 63, Sprecher - Sprachdaten - Sprachproduktion (3)

In Abb. 64 ist die Struktur der iterativen Komponente „Unterricht_Korrekturen“ zu sehen.

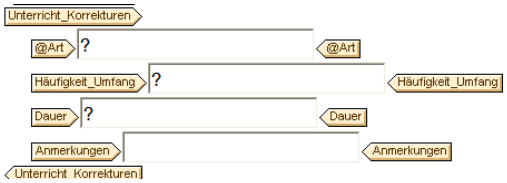

Abb. 64, Sprecher -Sprachdaten - Sprachproduktion (4)

Abb. 65 zeigt die im Erfassungsformular des Projekts „Deutsch heute“ für das Feld „Art“ im Komplex „Unterricht_Korrekturen“ bereitgestellten Werte. „Fremdkorrektur“ meint Korrektur des Sprechers durch andere Personen, „Selbstkorrektur“ heißt, dass der Sprecher seine sprachlichen Äußerungen gewöhnlich selbst korrigiert.

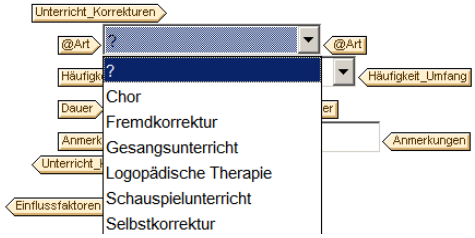

Abb. 65, Beispielwerte für Unterricht_Korrekturen

In das Feld „Sprachliche_Besonderheiten können produktionsrelevante Angaben wie z.B. „Stottern“, „Lispeln“, „Nuscheln“ etc. eingetragen werden.

Abb. 66, Sprecher - Sprachdaten - Sprachproduktion (5)

6.1.3.3. Sprachgebrauch

Abb. 67 zeigt die Struktur des fakultativen Abschnitts „Sprachgebrauch“. Eine wichtige Kategorie in diesem Abschnitt ist „Domäne“, worunter wir einen relevanten Kommunikationsbereich verstehen. Da i.d.R. Angaben über mehrere Kommunikationsbereiche zu erfassen sind, wurde der Abschnitt im Schema als iterativ gekennzeichnet. Als „Domäne“ gelten z.B. „Familie“, „Nachbarschaft“ und „Arbeitsplatz“.

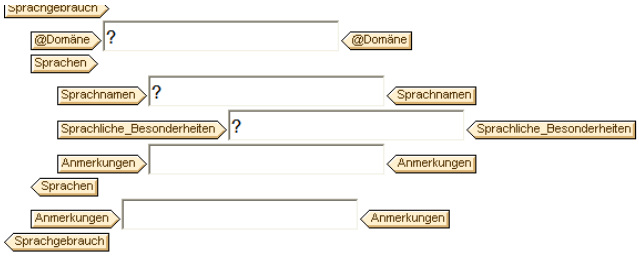

Abb. 67, Sprecher - Sprachdaten - Sprachgebrauch

Im Abschnitt „Sprachen“ werden zunächst die Namen der verwendeten Sprachen (z.B. Deutsch ; Türkisch“) erwartet. Unter dem Titel „Sprachliche_Besonderheiten“ kann man Hinweise auf weitere Aspekte, wie z.B. Dialektgebrauch, verzeichnen.

6.1.4. Beziehung zu anderem Sprecher

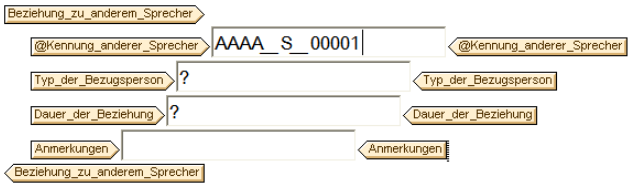

Abb. 68, Beziehung zu anderem Sprecher

In Abb. 68 sehen Sie die Struktur des fakultativen Bereichs, in dem über Beziehungen zwischen Sprechern informiert werden kann. Damit mehrere Beziehungen dokumentiert werden können, wurde der Abschnitt im Schema als iterativ gekennzeichnet.

An erster Stelle wird die Kennung des anderen Sprechers registriert, dann ist über den „Typ_der_Bezugsperson“ zu informieren. In diesem Feld könnte z.B. „Freund“, „Nachbar“, „Tochter“ etc. stehen. Für Angaben über die Dauer der Beziehung gibt es ein gleichlautendes Feld.

6.1.5. Sonstige Bezugspersonen

Neben dem Bereich der eigentlichen Sprecherdaten haben wir einen fakultativen Bereich für Daten über sonstige Bezugspersonen vorgesehen, der die fakultativen Teile „Bezugspersonen kompakt“ und „Einzelne Bezugsperson“ umfasst. Die Unterscheidung wurde angesichts der Notwendigkeit getroffen, verschiedenen Datenbeständen gerecht zu werden. Wir müssen zum einen kompakte Informationen über Gruppen von Bezugspersonen, zum anderen Daten für einzelne Bezugspersonen erfassen können.

6.1.5.1. Bezugspersonen kompakt

Wir gehen davon aus, dass sich auch bei kompakten Angaben über Gruppen von Bezugspersonen (z.B. „Eltern“, „Kinder“, „Freunde“, „Kollegen“) Personen- und Ortsdaten einerseits sowie Sprachdaten andererseits unterscheiden lassen. Falls der Gesamtkomplex gewählt wird, werden Personen- und Ortsdaten obligatorisch, Sprachdaten sind fakultativ, können also bei der Erstellung projektspezifischer Schemata übergangen werden.

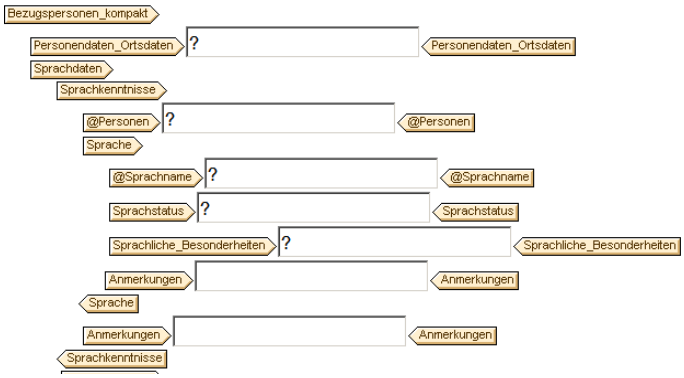

Abb. 69, Bezugspersonen kompakt (1)

An erster Stelle in Abb. 69 ist die Sammelkategorie „Personendaten_Ortsdaten“ zu sehen, die persönliche Angaben über Personengruppen und Informationen über sprachlich relevante Orte aufnehmen soll. Ein fiktives Beispiel dafür wäre: „Die Eltern sind seit 60 Jahren verheiratet und stammen beide aus Freiburg. Die drei Kinder leben und arbeiten in Frankfurt, Hamburg und München und besuchen die Herkunftsfamilie nur noch zu besonderen Anlässen.“

Sprachdaten gliedern wir hier in Sprachkenntnisse und Sprachgebrauch. Auf die Komponente „Sprachproduktion“, die für den Sprecher angelegt wurde, haben wir in diesem Bereich verzichtet, da wir davon ausgehen, dass diese Daten für Bezugspersonen nicht erhoben werden.

Die Struktur der Komponente „Sprachkenntnisse“ im Bereich „Bezugspersonen_kompakt“ ist ebenfalls in Abb. 69 dargestellt. Diese Komponente ist fakultativ und - für den Fall, dass Sprachkenntnisse mehrerer Personengruppen zu dokumentieren sind - iterativ. Im Unterschied zu Abschnitt 6.1.3.1., wo diese Komponente eingeführt wurde, haben wir hier auf den Teil „Kenntnisgrade“ verzichtet, da für Personengruppen vermutlich keine (einheitlichen) Kenntnisgrade zu ermitteln sind.

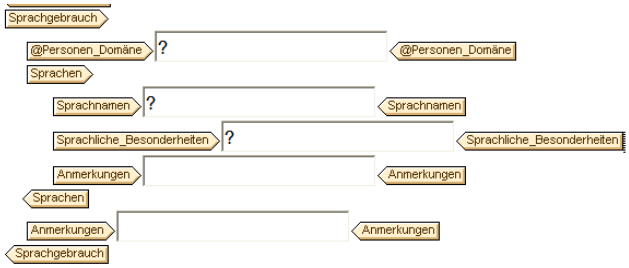

Abb. 70, Bezugspersonen kompakt (2)

An erster Stelle der Komponente „Sprachgebrauch“ steht das Feld „Personen_Domäne“, das Werte wie z.B. „Eltern untereinander“, „Eltern mit Kindern“ oder „Eltern mit Großeltern“ aufnehmen soll. Die übrigen Elemente wurden schon im Abschnitt 6.1.3.3. vorgestellt.

6.1.5.2. Einzelne Bezugsperson

Abb. 71, Einzelne Bezugsperson - Typ

Der fakultative Bereich „Einzelne_Bezugsperson“ umfasst das Feld „Typ“ (der Bezugsperson) für Werte wie z.B. „Mutter“ oder „Tante“ sowie Basisdaten, Ortsdaten und Sprachdaten. Das Feld „Typ“ ist obligatorisch, d.h. dass es berücksichtigt werden muss, wenn der Bereich „Ein-

zelne_Bezugsperson“ gewählt wird, Basisdaten, Ortsdaten und Sprachdaten sind fakultativ, können aber nicht alle bei der Erstellung eines projektspezifischen Schemas ausgeschlossen werden.

6.1.5.2.1. Basisdaten

Besonderheiten der fakultativen Basisdaten für einzelne Bezugspersonen sind am Anfang von Abb. 72 dargestellt. Dort finden Sie die Felder „Status_der_Bezugsperson“ (für den Sprecher), wofür die Werte „Relevant“ und „Peripher“ vorgesehenen sind, sowie „Dauer der Beziehung“. Zu allen anderen Feldern in diesem Komplex gibt es Entsprechungen in den Sprecher-Basisdaten, die in Abschnitt 6.1.1. beschrieben wurden.

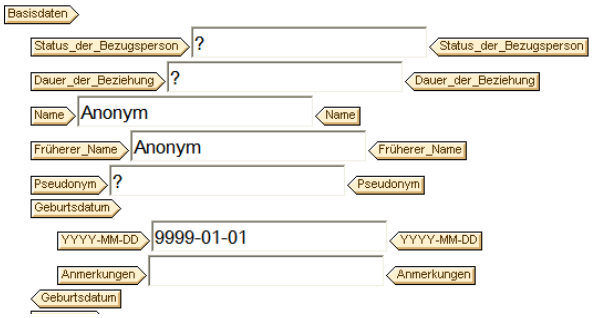

Abb. 72, Einzelne Bezugsperson - Basisdaten (1)

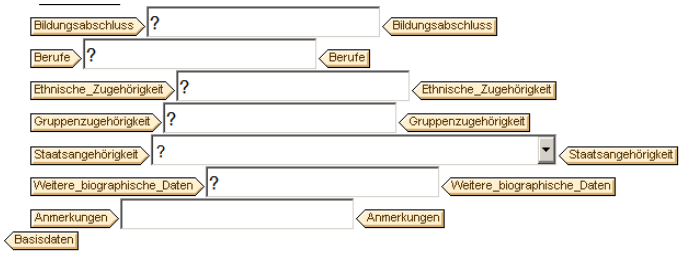

Abb. 73, Einzelne Bezugsperson - Basisdaten (2)

6.1.5.2.2. Ortsdaten

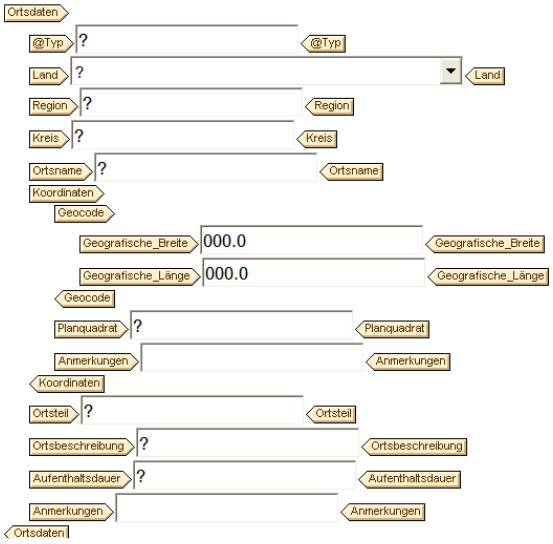

Abb. 74, Einzelne Bezugsperson - Ortsdaten

In Abb. 74 sehen Sie die Struktur der fakultativen Ortsdaten für einzelne Bezugspersonen. Sie stimmt mit der der Ortsdaten für Sprecher überein, über die Sie sich in Abschnitt 6.1.2. informieren können.

6.1.5.2.3. Sprachdaten

Der fakultative Komplex „Sprachdaten“ im Bereich „Einzelne Bezugspersonen“ umfasst die fakultativen Komponenten „Sprachkenntnisse“ und „Sprachgebrauch“. Auf die Komponente „Sprachproduktion“, die für den Sprecher angelegt wurde, haben wir in diesem Bereich verzichtet, da wir davon ausgehen, dass diese Daten für Bezugspersonen nicht erhoben werden.

6.1.5.2.3.1. Sprachkenntnisse

Für Informationen über Sprachkenntnisse von Sprechern und einzelnen Bezugspersonen haben wir eine einheitliche Struktur gewählt. Erläuterungen dazu finden Sie in Abschnitt 6.1.3.1.

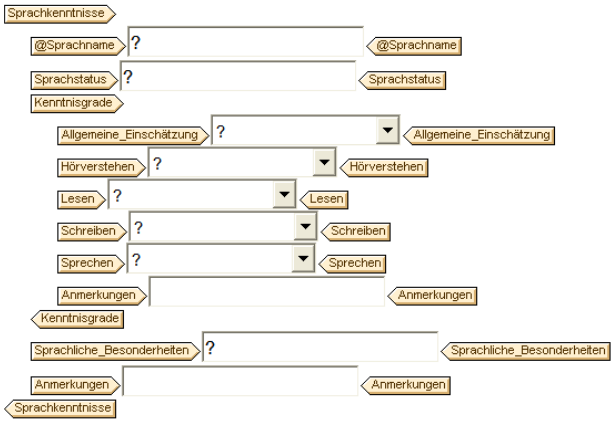

Abb. 75, Einzelne Bezugsperson - Sprachdaten - Sprachkenntnisse

6.1.5.2.3.2. Sprachgebrauch

Auch Informationen über den Sprachgebrauch von Sprechern und einzelnen Bezugspersonen werden in einer einheitlichen Struktur erfasst. Erläuterungen dazu gibt es in Abschnitt 6.1.3.3.

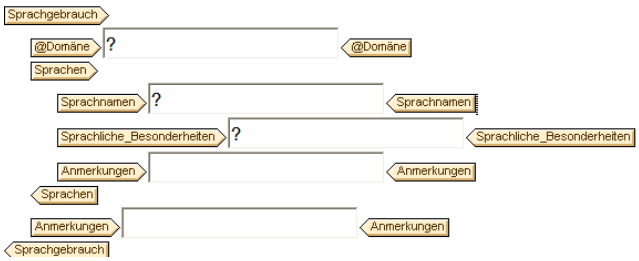

Abb. 76, Einzelne Bezugsperson - Sprachdaten - Sprachgebrauch

6.2. Rechteverwaltung

Im Komplex „Rechteverwaltung“ sollen rechtliche Aspekte der Datenerhebung sowie rechtsrelevante Vereinbarungen mit Sprechern und ggf. auch Bezugspersonen über Schutz und Verwendung ihrer Daten dokumentiert werden. Im Folgenden unterscheiden wir zwischen personenbe-

zogenen Daten und Korpusbestandteilen, wobei Korpusbestandteile personenbezogene Daten enthalten können.

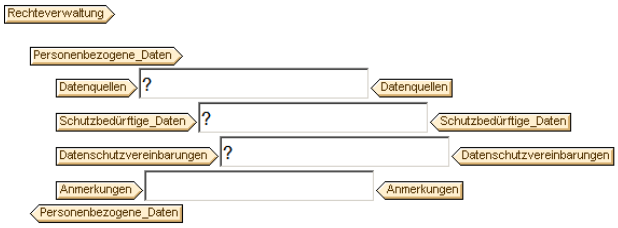

Abb. 77, Sprecher - Rechteverwaltung - Personenbezogene Daten

Zu den rechtsrelevanten Aspekten zählt die Frage, aus welchen Quellen die erhobenen personenbezogenen Daten stammen. In der Regel stammen sie von den Sprechern, es können aber auch Bezugspersonen befragt oder schriftliche Quellen ausgewertet worden sein. Im Feld „Schutzbedürftige_Daten“ kann man festhalten, ob nur besondere Arten oder alle personenbezogener Daten zu schützen sind. Mit „Datenschutzvereinbarungen“ meinen wir Vereinbarungen mit Sprechern und Bezugspersonen über den Schutz der im vorigen Feld genannten Daten. Solche Vereinbarungen können vorsehen, dass diese Daten nur im Rahmen des erhebenden Projekts verwendet und danach gelöscht werden oder auf bestimmten Wegen für bestimmte Zwecke an Dritte weitergegeben werden dürfen.

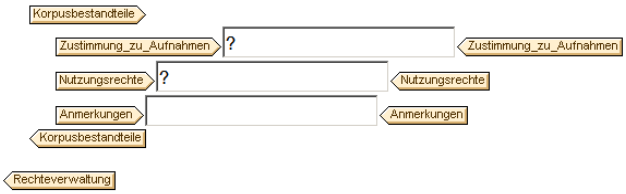

Abb. 78, Sprecher - Rechteverwaltung - Korpusbestandteile

Die Zustimmung der Sprecher zu den Aufnahmen ist eine wesentliche Voraussetzung für die Verwendung von Aufnahmen und Transkripten. Im entsprechenden Feld sollte man daher festhalten, ob und wann die Sprecher über den Zweck der Aufnahmen informiert wurden und in welcher Form - schriftlich oder (aus welchen Gründen?) mündlich - sie den Aufnahmen zugestimmt haben. Schließlich werden Nutzungsrechte an Korpusbestandteilen dokumentiert, die von der wissenschaftlichen Auswertung im Daten erhebenden Projekt bis zur Veröffentlichung im Internet reichen können.

6.3. Zusatzmaterial

Unter Zusatzmaterial auf Sprecherebene verstehen wir Dokumente wie z.B. Sprecherfotos oder schriftliche Vereinbarungen mit den Sprechern, die als Korpusbestandteile gelten können. Die Struktur für die Dokumentation von Zusatzmaterial auf Sprecherebene stimmt mit der für Zusatzmaterial auf Ereignis- und auf Sprechereignisebene (vgl. Abschnitte 5.4. und 5.8.) überein. Der Komplex ist fakultativ. Da pro Sprecher mehrere zusätzliche Dokumente vorliegen können, wurde er im Schema als iterativ gekennzeichnet.

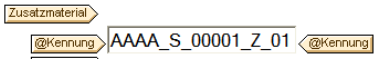

Abb. 79, Sprecher - Zusatzmaterial - Kennung

Die Kennung des Zusatzmaterials ist ebenenspezifisch und umfasst hier die Sprecherkennung, den Kennbuchstaben Z (für „Zusatzmaterial“) und eine zweistellige laufende Nummer. Ein Beispiel finden Sie in Abb. 79.

6.3.1. Basisdaten

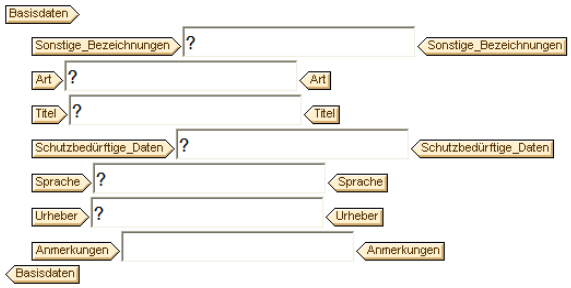

Abb. 80, Sprecher - Zusatzmaterial – Basisdaten

An erster Stelle der Komponente „Basisdaten“ können Bezeichnungen eingetragen werden, die vor der Kennung vergeben wurden. Zusatzmaterialien können Daten enthalten, die nach dem Willen der Urheber und aus datenschutzrechtlichen Gründen Außenstehenden nicht kenntlich werden dürfen, wie z.B. persönliche Sprecherdaten. Für entsprechende Informationen wurde das Feld „Schutzbedürftige_Daten“ bereitgestellt. „Urheber“ steht für Autoren, Grafiker, Fotografen etc. Die übrigen Felder in diesem Komplex müssen vermutlich nicht erläutert werden.

6.3.2. Technische Fassungen

Da zusätzliche Dokumente in verschiedenen technischen Fassungen vorliegen können, haben wir die fakultativen und iterativen Komponenten „Analoge_Fassung“ und „Digitale_Fassung“ vorgesehen. Wenigstens eine Komponente muss bei der Erstellung eines projektspezifischen Schemas gewählt werden.

Zunächst wird eine Kennung abgefragt, die aus der Kennung des Zusatzmaterials, dem Kürzel AF (für „Analoge_Fassung“) bzw. DF (für „Digitale_Fassung“) und einer zweistelligen laufenden Nummer besteht. Beispiele finden Sie in den Abb. 81 und 82.

„Datenschutz“ meint technische Maßnahmen zum Datenschutz, wie z.B. die Maskierung von Personennamen in Texten.

Da eine technische Fassung auf mehreren Datenträgern gespeichert sein kann, wurde dieser Abschnitt im Schema als iterativ gekennzeichnet. An erster Stelle dieses Abschnitts wird eine eindeutige Inventarnummer des zu dokumentierenden Datenträgers erwartet. Im nächsten Schritt ist über den Typ des Datenträgers (z.B. Papier oder Mikrofilm) zu informieren.

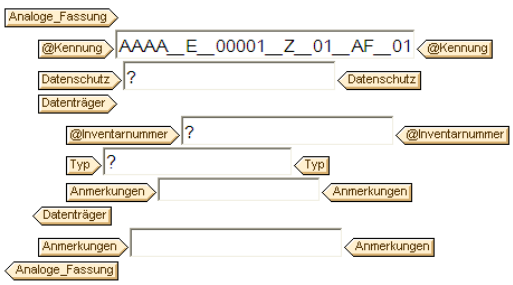

Abb. 81, Sprecher - Zusatzmaterial - Analoge Fassung

Nur für digitale Fassungen relevant sind die Felder „Typ“, „Digitalisierungssoftware“ und „Elektronischer_Speicherort“ sowie die Komponente „Technische_Daten“.

Grundlage für eine Typisierung digitaler Fassungen von Zusatzmaterial sind v.a. technische Daten. Als Typenbezeichnungen dient das Kürzel DFT (digitale Fassung) in Verbindung mit einer zweistelligen Nummer.

Die Dateigröße soll in Bytes angegeben werden. Im Feld „Digitalisierungssoftware“ kann das Programm, mit dem eine analoge Fassung digitalisiert wurde, genannt werden. Im Feld „Elektronischer_Speicherort“ wird eine URL oder ein Pfadname erwartet.

Das erste Feld der Komponente „Technische_Daten“ ist für eine Information über das Dateiformat vorgesehen. „Character_Encoding“ steht für die Zeichencodierung in einer Textdatei (z.B. ASCII oder UTF-16BE).

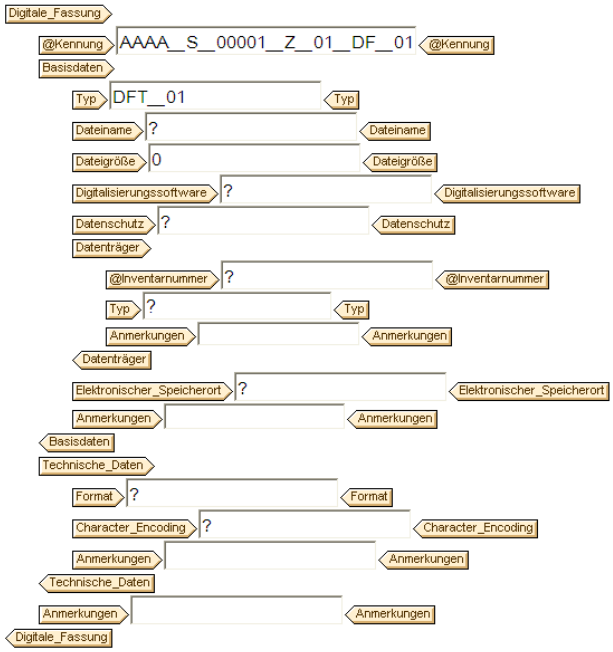

Abb. 82, Sprecher - Zusatzmaterial - Digitale Fassung

6.3.3. Archivierung und Distribution

In den Abb. 83 und 84 werden die Bausteine „Archivierung“ und „Distribution“ vorgestellt, die Sie möglicherweise schon aus der Beschreibung des Ereignisschemas kennen. Sie sollen Informationen über rechtliche und organisatorische Aspekte der Korpusbestandteile aufnehmen.

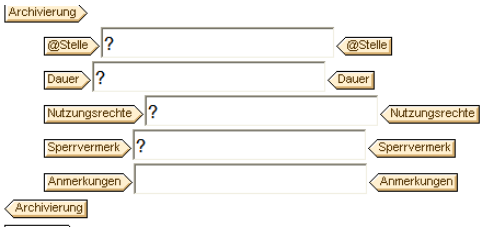

Abb. 83, Sprecher - Zusatzmaterial - Archivierung

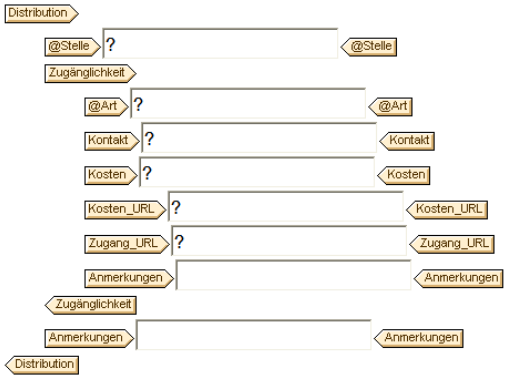

Abb. 84, Sprecher - Zusatzmaterial - Distribution

Das Modul „Archivierung“ wurde im Schema als iterativ gekennzeichnet. Zunächst soll der Name der archivierenden Stelle vermerkt werden. Es folgt ein Feld für Informationen über die vorgesehene Archivierungsdauer. Hier könnte z.B. „Bis 2018“ oder „Langfristig“ stehen. Die Nutzungsrechte der archivierenden Stelle können von der ausschließlichen Nutzung durch den Projektleiter bis hin zur Veröffentlichung eines Dokuments im Internet reichen. Sperrvermerke, wie z.B. „Bis 2010 für Externe gesperrt“, können die Nutzungsmöglichkeiten einschränken.

Das Modul „Distribution“ ist ebenfalls iterativ und umfasst neben einem Feld für den Namen der für die Distribution zuständigen Stelle die iterative Komponente „Zugänglichkeit“. In dieser Komponente sollen folgende Angaben verzeichnet werden: Art der Zugänglichkeit, E-Mail-Kontaktadresse, Angaben über die Kosten, ggf. eine URL dieser Angaben sowie ggf. eine URL, die den direkten Zugang zum jeweiligen Korpusbestandteil ermöglicht.

6.4. Dokumentationsgeschichte

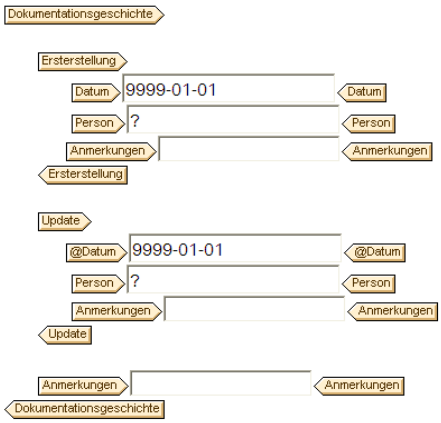

Abb. 85, Sprecher - Dokumentationsgeschichte

Das Schema für die Erfassung allgemeiner Sprecherdaten enthält wie das Schema für die Dokumentation von Korpusbestandteilen auf Ereignisebene den Komplex „Dokumentationsgeschichte“.

7. Generisches Schema für die Dokumentation von Zusatzmaterial auf der Korpus ebene

Unter „Zusatzmaterial auf der Korpusebene“ verstehen wir Dokumente, die zusätzlich zu Aufnahmen, Transkripten und Zusatzmaterialien auf Ereignis-, Sprechereignis- und Sprecherebene

vorhanden sein können. Das sind z.B. Transkriptionskonventionen, ein Interviewleitfaden, Wortlisten, verschiedene Varianten der Wenkersätze, ggf. auch Spezifikationen für die Validierung von Korpusdaten sowie Dokumente, die die Ergebnisse solcher Qualitätsprüfungen enthalten [17].

Die Struktur für die Dokumentation von Zusatzmaterial auf Korpusebene stimmt mit der für Zusatzmaterial auf Ereignis-, Sprechereignis- und Sprecherebene überein. Lediglich die Kennungen sind ebenenspezifisch. Wenn Sie den Komplex „Zusatzmaterial“ des Ereignisschemas oder des Schemas für allgemeine Sprecherdaten kennengelernt haben, können Sie weite Teile der folgenden Darstellung übergehen.

Das Schema enthält obligatorische und fakultative Komponenten. Obligatorische Komponenten sind in allen projektspezifischen Subschemata zu berücksichtigen, fakultative Komponenten stehen zur Wahl und müssen in den Subschemata nicht verwendet werden. Wenn man sie verwendet, sind alle Kennungsfelder und die mit ? gekennzeichneten Felder zu bearbeiten.

Eingaben für fehlende Daten in diesen Feldern sind standardisiert: „Nicht dokumentiert“ bedeutet: Es kann ein Datum geben, das bei der Datenerfassung jedoch nicht bekannt ist. Ein Beispiel dafür wäre: „Urheber: Nicht dokumentiert“ - zu lesen als: „Der Name des Urhebers ist nicht dokumentiert.“ „Nicht vorhanden“ bedeutet: Es gibt kein Datum. Ein Beispiel dafür wäre: „Schutzbedürftige Daten: Nicht vorhanden“ - zu lesen als: „Es gibt in diesem Dokument keine schutzbedürftigen Daten.“

Das an vielen Stellen vorgesehene Feld „Anmerkungen“ ist für Anmerkungen zu Angaben in anderen Feldern und für nicht kategorisierte Angaben vorgesehen. Das Feld kann leer bleiben.

Einzelne Komponenten des Schemas wurden als iterativ gekennzeichnet, d.h. dass sie bei der Datenerfassung vervielfältigt werden können.

Die nachfolgenden Abbildungen stammen aus einem projektneutralen Erfassungsformular, das zu Demonstrationszwecken angelegt wurde.

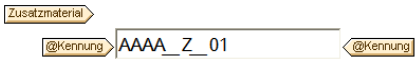

Abb. 86, Zusatzmaterial - Kennung

Zuerst wird eine Kennung für Zusatzmaterial auf Korpusebene generiert. Diese Kennung setzt sich zusammen aus der Korpuskennung, dem Kennbuchstaben Z (für „Zusatzmaterial“) und einer zweistelligen laufenden Nummer. Ein Beispiel finden Sie in Abb. 86.

7.1. Basisdaten

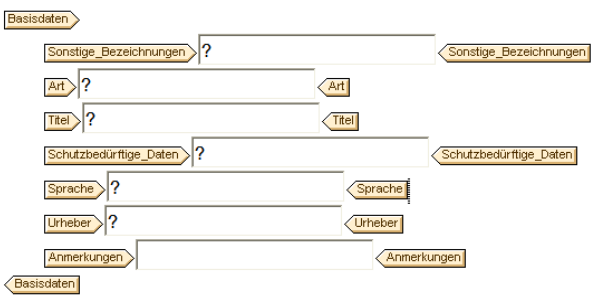

Abb. 87, Zusatzmaterial - Basisdaten

An erster Stelle der Komponente „Basisdaten“ können Bezeichnungen eingetragen werden, die vor der Kennung vergeben wurden. Zusatzmaterialien können Daten enthalten, die nach dem Willen der Urheber und aus datenschutzrechtlichen Gründen Außenstehenden nicht kenntlich werden dürfen, wie z.B. persönliche Sprecherdaten. Für entsprechende Informationen wurde das Feld „Schutzbedürftige_Daten“ bereitgestellt. „Urheber“ steht für Autoren, Grafiker, Fotografen etc. Die übrigen Felder in diesem Komplex müssen vermutlich nicht erläutert werden.

7.2. Technische Fassungen

Da zusätzliche Dokumente in verschiedenen technischen Fassungen vorliegen können, haben wir die fakultativen und iterativen Komponenten „Analoge_Fassung“ und „Digitale_Fassung“ vorgesehen. Wenigstens eine Komponente muss bei der Erstellung eines projektspezifischen Schemas gewählt werden.

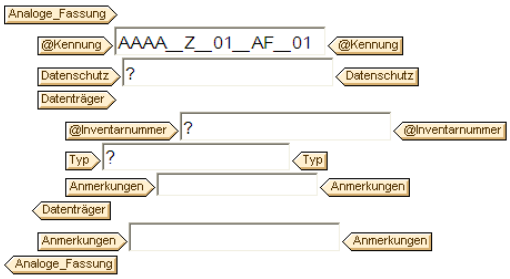

Abb. 88, Zusatzmaterial - Analoge Fassung

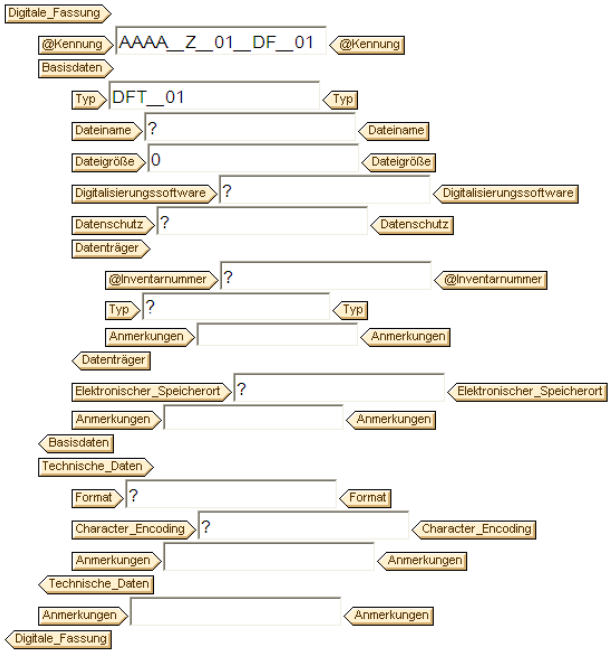

Abb. 89, Zusatzmaterial - Digitale Fassung

Zunächst wird eine Kennung abgefragt, die aus der Kennung des Zusatzmaterials, dem Kürzel AF (für „Analoge_Fassung“) bzw. DF (für „Digitale_Fassung“) und einer zweistelligen laufenden Nummer besteht. Beispiele finden Sie in den Abb. 88 und 89. „Datenschutz“ meint technische Maßnahmen zum Datenschutz, wie z.B. die Maskierung von Personennamen in Texten.

Da eine technische Fassung auf mehreren Datenträgern gespeichert sein kann, wurde dieser Abschnitt im Schema als iterativ gekennzeichnet. An erster Stelle dieses Abschnitts wird eine

eindeutige Inventarnummer des zu dokumentierenden Datenträgers erwartet. Im nächsten Schritt ist über den Typ des Datenträgers (z.B. Papier oder Mikrofilm) zu informieren.

Nur für digitale Fassungen relevant sind die Felder „Typ“, „Digitalisierungssoftware“ und „Elektronischer_Speicherort“ sowie die Komponente „Technische_Daten“.

Grundlage für eine Typisierung digitaler Fassungen von Zusatzmaterial sind v.a. technische Daten. Als Typenbezeichnungen dient das Kürzel DFT (digitale Fassung) in Verbindung mit einer zweistelligen Nummer.

Die Dateigröße soll in Bytes angegeben werden. Im Feld „Digitalisierungssoftware“ kann das Programm, mit dem eine analoge Fassung digitalisiert wurde, genannt werden. Im Feld „Elektronische Speicherort“ wird eine URL oder ein Pfadname erwartet.

Das erste Feld der Komponente „Technische_Daten“ ist für eine Information über das Dateiformat vorgesehen. „Character_Encoding“ steht für die Zeichencodierung in einer Textdatei (z.B. ASCII oder UTF-16BE).

7.3. Archivierung und Distribution

Auch dieses Schema enthält die Module „Archivierung“ und „Dokumentation“.

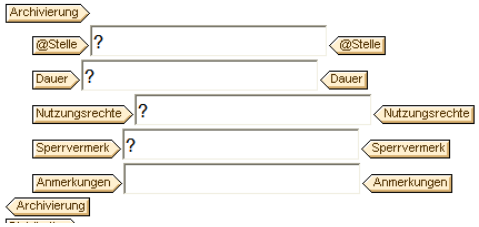

Abb. 90, Zusatzmaterial - Archivierung

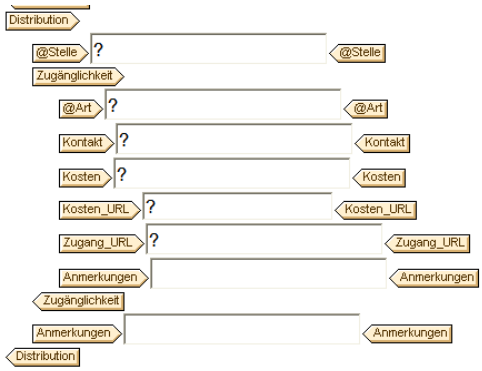

Abb. 91, Zusatzmaterial - Distribution

„Archivierung“ ist iterativ, kann also bei der korpusspezifischen Datenerfassung vervielfältigt werden. Zunächst soll der Name der archivierenden Stelle vermerkt werden. Es folgt ein Feld für Informationen über die vorgesehene Archivierungsdauer. Hier könnte z.B. „Bis 2018“ oder „Langfristig“ stehen. Die Nutzungsrechte der archivierenden Stelle können von der ausschließlichen Nutzung durch den Projektleiter bis hin zur Veröffentlichung eines Dokuments im Internet reichen. Sperrvermerke, wie z.B. „Bis 2010 für Externe gesperrt“, können die Nutzungsmöglichkeiten einschränken.

Das Modul „Distribution“ ist ebenfalls iterativ und umfasst neben einem Feld für den Namen der für die Distribution zuständigen Stelle die iterative Komponente „Zugänglichkeit“, wo folgende Angaben verzeichnet werden können: Art der Zugänglichkeit, E-Mail-Kontaktadresse, Angaben

über die Kosten, ggf. eine URL dieser Angaben sowie ggf. eine URL, die den direkten Zugang zum jeweiligen Korpusbestandteil ermöglicht.

7.4. Dokumentationsgeschichte

Das Schema für die Dokumentation von Zusatzmaterial auf der Korpusebene umfasst den in allen Schemata enthaltenen Komplex „Dokumentationsgeschichte“.

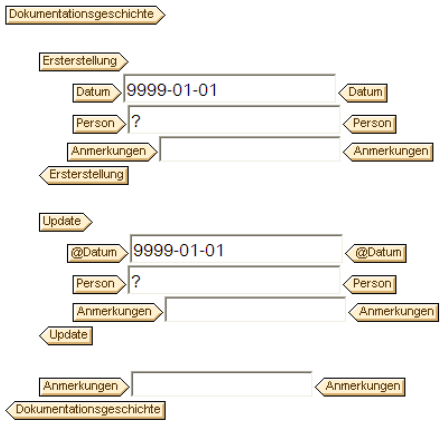

Abb. 92, Zusatzmaterial - Dokumentationsgeschichte

8. Generisches Schema für die Korpusbeschreibung

Der vierte Bereich in unserem Datenmodell, der mithilfe eines (XML-)Schemas gestaltet wurde, ist die Korpusbeschreibung, die einen systematischen Überblick über Erstellung, Zusammensetzung, Bearbeitungsstand und Verwaltung eines Korpus ermöglichen soll. Bei der Gestaltung dieses Schemas haben wir auf Konzepte und Strukturelemente zurückgegriffen, die im 5. Abschnitt des vorliegenden Textes beschrieben sind

Auch dieses Schema enthält obligatorische und fakultative Komponenten. Obligatorische Komponenten sind in allen projektspezifischen Subschemata zu berücksichtigen, fakultative Komponenten stehen zur Wahl. Wenn man sie nutzt, sind alle Kennungsfelder und die mit ? gekennzeichneten Felder zu bearbeiten.

Eingaben für fehlende Daten in diesen Feldern sind standardisiert: „Nicht dokumentiert“ bedeutet: Es kann ein Datum geben, das bei der Datenerfassung jedoch nicht bekannt ist. Ein Beispiel dafür wäre: „Laufzeit: Nicht dokumentiert“ - zu lesen als: „Die Laufzeit des Erstellungsprojekt ist nicht dokumentiert.“ „Nicht vorhanden“ bedeutet: Es gibt kein Datum. Ein Beispiel dafür wäre: „Beschreibung: Nicht vorhanden“ - zu lesen als: „Es gibt keine Beschreibung des Erstellungsprojekt bzw. der Korpusarbeiten.“ In ein Feld („Datenrate“) kann auch der Wert „Nicht relevant“ eingegeben werden.

Das an vielen Stellen vorgesehene Feld „Anmerkungen“ ist für Anmerkungen zu Angaben in anderen Feldern und für nicht kategorisierte Angaben vorgesehen. Das Feld kann leer bleiben.

Einzelne Komponenten des Schemas wurden als iterativ gekennzeichnet, d.h. dass sie bei der Datenerfassung vervielfältigt werden können.

Die nachfolgenden Abbildungen stammen aus einem projektneutralen Erfassungsformular, das zu Demonstrationszwecken angelegt wurde.

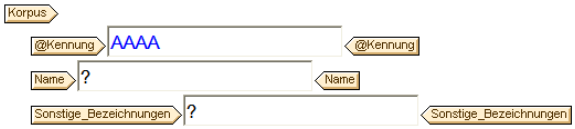

Abb. 93, Korpusbeschreibung (1)

Jede Korpusbeschreibung beginnt mit einer vierstelligen Korpuskennung, dem Korpusnamen und möglichen früheren Bezeichnungen.

8.1. Erstellungsprojekt

Unter „Erstellungsprojekt“ verstehen wir das Projekt, das ein Korpus aufgebaut hat. Bei der Korpuserstellung können Materialien aus anderen Projekten verwendet worden sein, was in bestimmten Komponenten des Komplexes „Korpusbestandteile“ (vgl. 8.3.) vermerkt werden kann.

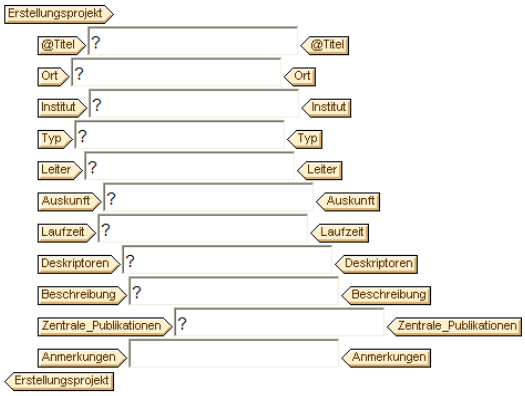

Abb. 94, Korpusbeschreibung – Erstellungsprojekt

Der in Abb. 94 gezeigte Abschnitt „Erstellungsprojekt“ wurde im Schema als iterativ gekennzeichnet, um Projektkooperationen bei der Korpuserstellung dokumentieren zu können. Er ist an der „IDS-Dokumentation zur Germanistischen Sprachwissenschaft - Sprachwissenschaftliche Forschungsvorhaben“ [18] orientiert, deren Struktur wir für unsere Zwecke erweitert haben.

8.2. Aufzeichnungsobjekte

In diesem Komplex greifen wir die Konzepte „Ereignis“ und „Sprechereignis“ wieder auf, die in den Abschnitten 5.1. und 5.5. eingeführt wurden. Zur Erinnerung wiederholen wir im Folgenden die o.g. Definitionen.

Unter „Ereignis“ (E) verstehen wir eine Phase des sozialen Geschehens, die von Beteiligten bzw. Korpusproduzenten als abgrenzbare Einheit wahrgenommen und aufgezeichnet wird. Unter „Sprechereignis“ (SE) verstehen wir den aufgezeichneten kommunikativen Gehalt eines Ereignisses bzw. Segmente dieses Gehalts. Diese Definitionen sind bewusst sehr allgemein gehalten. Wir stellen lediglich für die Dokumentation von Korpusbestandteilen relevante Konzepte bereit, keine linguistischen Segmentierungskriterien.

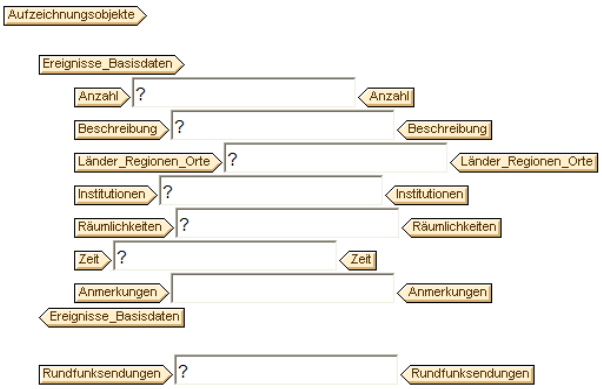

Abb. 95, Korpusbeschreibung - Aufzeichnungsobjekte – Ereignis-Basisdaten, Rundfunksendungen

Im Feld „Anzahl“ der Ereignis_Basisdaten kann man die Anzahl der aufgezeichneten Ereignisse notieren. Im Feld „Beschreibung“ wird eine kurze inhaltliche Charakterisierung der Ereignisse erwartet.

Im Anschluss an die Ortsangaben sollen die gesellschaftlichen Kontexte der Ereignisse charakterisiert werden. Dafür ist das Feld „Institutionen“ vorgesehen. Wir verstehen „Institution“ i.S.v. „Organisation“

Im Feld Räumlichkeiten kann man über das räumliche Umfeld der Ereignisse berichten. „Zeit“ steht für den Zeitraum, in dem die dokumentierten Ereignisse stattgefunden haben. Das Feld „Rundfunksendungen“ wurde für Korpora mit Mitschnitten solcher Sendungen eingerichtet. Hier sollten ggf. Anzahl und Typen der Sendungen (Hörfunksendung, Fernsehsendung) eingetragen werden.

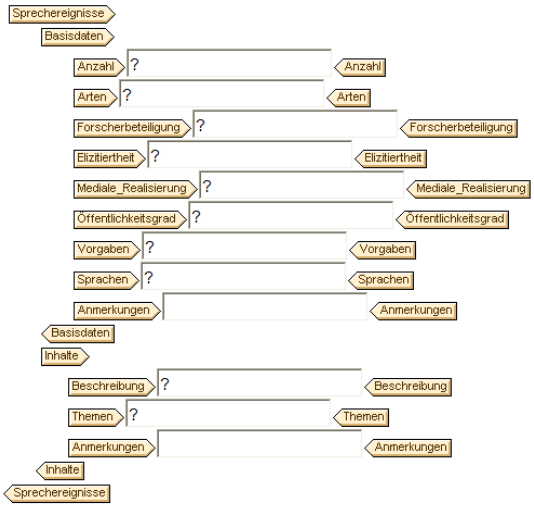

Abb. 96, Korpusbeschreibung - Aufzeichnungsobjekte - Sprechereignisse

Wir verwenden „Art“ in der Beschreibung der aufgezeichneten Sprechereignisse anstelle von Kategorien wie „Textsorte“, „Texttyp“, „Interaktionstyp“, „Gesprächstyp“, „Diskurstyp“, „Genre“, „Gattung“, die aus verschiedenen Forschungsansätzen stammen, um Daten aus allen Bereichen aufnehmen zu können. Wir denken dabei an Werte wie „Erzählung“, „Rede“, „Anleitung“, „Beschreibung“, „Benennung“, „Übersetzung“, „Interview“, „Beratung“, „Diskussion“, „Begrüßung“ etc. Mit diesen Beispielwerten wollen wir keine Vorentscheidung über eine im Einzelfall anzuwendende Systematik treffen.

Für „Forscherbeteiligung“ haben wir die Werte „Verbal beteiligt“, „Nicht verbal beteiligt“ und „Nicht vorhanden“ (für „Forscher nicht anwesend“) vorgesehen. „Elizitierung“ ist eine Technik zur Erhebung sprachlicher Daten, bei der die Informanten systematisch zu Äußerungen veranlasst werden. Wir haben die Werte „Elizitiert“ und „Nicht elizitiert“ vorgesehen.

„Mediale_Realisierung“ steht für den Kommunikationskanal (wie z.B. „Face to Face“, „Telefon“, „Hörfunk“). Für das Feld „Öffentlichkeitsgrad“ werden die Werte „Öffentlich“ und „Nicht öffentlich“ bereitgestellt Über Instruktionen von Sprechern durch Aufnahmeleiter und ggf. auch über Materialien, die den Sprechern zur Lösung bestimmter Aufgaben vorgelegt wurden, kann man im Feld „Vorgaben“ informieren. Im Feld „Sprachen“ sind die in den Sprechereignissen verwendeten Sprachen zu verzeichnen.

Im Komplex „Inhalte“ ist ein Feld für eine Beschreibung der Sprechereignisse vorgesehen. Themenangaben sollten stichwortartig sein (z.B. „Lebenslauf“, „berufliche Aufgaben“, “sprachliche Entwicklung“).

Abb. 97, Korpusbeschreibung - Aufzeichnungsobjekte - Sprecher

Im Unterschied zum generischen Schema für die Dokumentation von Korpusbestandteilen auf Ereignisebene und zum generischen Schema für allgemeine Sprecherdaten wurden hier nur zwei Felder für Sprecherdaten vorgesehen. Im ersten Feld soll die Anzahl der Sprecher verzeichnet werden, im zweiten Feld kann die Sprecherauswahl beschrieben sowie zusammenfassend über Sprechermerkmale und sprachliche Besonderheiten informiert werden.

8.3. Korpusbestandteile

Zu den Korpusbestandteilen zählen wir Quellaufnahmen von Ereignissen, sprechereignisspezifische Aufnahmen, Transkripte und Zusatzmaterial auf Ereignis-, Sprechereignis-, Sprecher- und Korpusebene. Die Strukturen der entsprechenden Komplexe in der Korpusbeschreibung unterscheiden sich kaum von den entsprechenden Strukturen in den oben beschriebenen Schemata. Das ermöglicht in der Korpusbeschreibung eine Bilanzierung des jeweiligen Bestandes aufgrund der für die einzelnen Bezugsobjekte gesammelten Daten.

8.3.1. Quellaufnahmen

Unter „Quellaufnahmen“ verstehen wir Rohdaten, Originalaufnahmen von Ereignissen oder Aufnahmen, die für die dokumentierende Stelle Originalcharakter haben. Diese Aufnahmen können Quellen für sprechereignisspezifische Kopien sein. Da nicht jedes Korpus Quellaufnahmen umfasst und da Quellaufnahmen unterschiedlichen Typs vorliegen können, wurde der Komplex im Schema als fakultativ und iterativ gekennzeichnet.

Abb. 98, Korpusbestandteile - Quellaufnahmen (1)

Der Typ der Aufnahmen (Audio, Video und ggf. Tonkopie von Video) sollte im gleichnamigen Feld notiert werden.

8.3.1.1. Basisdaten

Die Anzahl der Quellaufnahmen des genannten Typs kann man im gleichnamigen Feld verzeichnen. Im Anschluss daran wird dokumentiert, ob das Korpus vollständige und / oder unvollständige Aufnahmen der Ereignisse umfasst. Im nächsten Schritt sollte man Angaben über die Dauer der einzelnen Aufnahmen (z.B. „zwischen 4 und 50 Minuten“) und die Gesamtdauer der Quellaufnahmen des genannten Typs notieren.

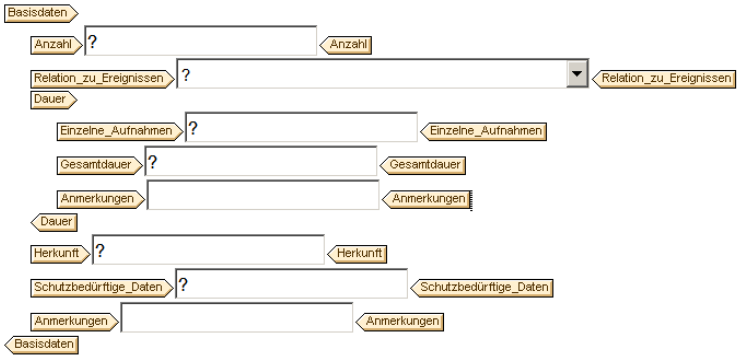

Abb. 99, Korpusbestandteile - Quellaufnahmen - Basisdaten

Quellaufnahmen müssen nicht unbedingt aus dem dokumentierten Erstellungsprojekt stammen. Sie können aus anderen Projekten bzw. Korpora übernommen worden sein. Um solchen Fällen gerecht werden zu können, wurde das Feld „Herkunft“ in die Basisdaten eingefügt. Aufnahmen können Daten enthalten, die nach dem Willen der Urheber und aus datenschutzrechtlichen Gründen Außenstehenden nicht kenntlich werden dürfen, wie z.B. persönliche Sprecherdaten. Für entsprechende Informationen wurde das Feld „Schutzbedürftige_Daten“ bereitgestellt.

8.3.1.2. Aufnahmetechnik

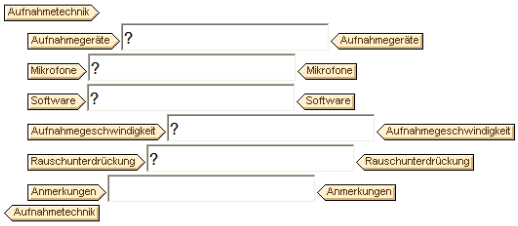

Abb. 100, Korpusbestandteile - Quellaufnahmen - Aufnahmetechnik

Unter dem Stichwort „Aufnahmetechnik“ werden Informationen über die Aufnahmeapparatur (Aufnahmegerät, Mikrofone), eine ggf. eingesetzte Aufzeichnungssoftware, die Aufnahmegeschwindigkeit (bei Spulentonbandaufnahmen relevante Angabe in cm/s) und Rauschunterdrückungsverfahren (z.B. Dolby B) zusammengefasst.

8.3.1.3. Technische Fassungen

Quellaufnahmen liegen in bestimmten technischen Fassungen vor. Das können analoge und / oder digitale Fassungen sein. Für jeden Typ gibt es einen eigenen Abschnitt im Schema. Beide

Abschnitte sind fakultativ und iterativ, wenigsten einen Abschnitt muss bei der Erstellung projektspezifischer Schemata übernommen werden.

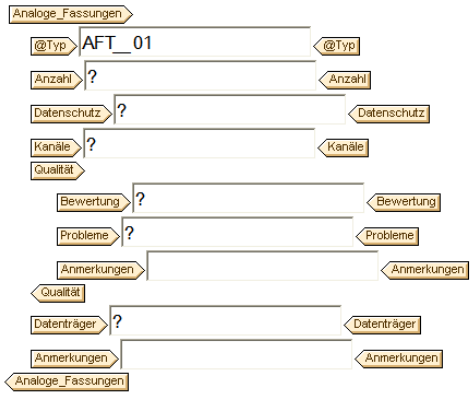

Abb. 101, Korpusbestandteile - Quellaufnahmen - Analoge Fassungen

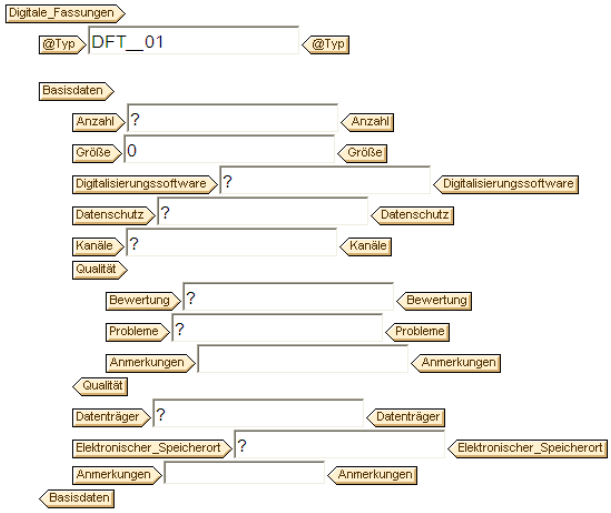

Abb. 102, Korpusbestandteile - Quellaufnahmen - Digitale Fassungen (1)

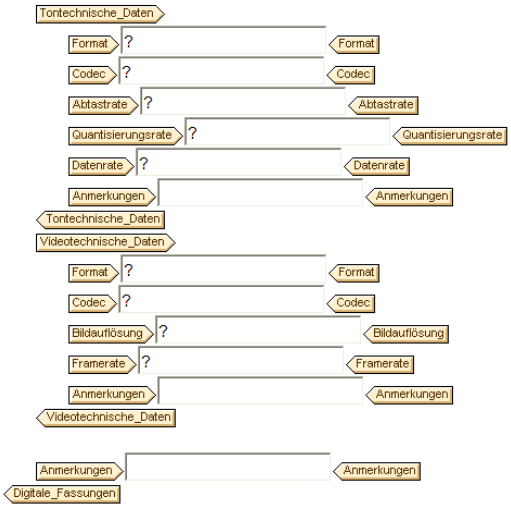

Abb. 103, Korpusbestandteile - Quellaufnahmen - Digitale Fassungen (2)

In allen Abschnitten werden Typen definiert. Bei analogen Fassungen sind die Werte in den Feldern „Datenschutz“ und „Kanäle“ für eine Typisierung relevant, bei digitalen Fassungen auch Informationen über die technischen Daten (z.B. das Dateiformat). Die Typenbezeichnungen

setzen sich zusammen aus „AFT“ (für „Analoge_Fassungen_Typ“) bzw. „DFT“ (für „Digiale_Fassungen_Typ“) und einer zweistelligen Nummer. Diese Typenbezeichnungen werden auch bei der Dokumentation einzelner Aufnahmen verwendet.

Im Feld „Anzahl“ soll über die Anzahl der technischen Fassungen des jeweiligen Typs informiert werden. „Datenschutz“ meint technische Maßnahmen zum Datenschutz, wie z.B. die Anonymisierung von Personennamen durch Verzerrung. Das Feld „Kanäle“ steht für Angaben wie „Mono“ oder „Stereo“ bereit. Für Angaben über die Qualität der Fassungen wurden zwei Felder bereitgestellt: „Bewertung“ und „Probleme“. Über die Datenträger kann im gleichnamigen Feld informiert werden.

8.3.1.4. Archivierung und Distribution

Auch die Module „Archivierung“ und „Distribution“, die Sie möglicherweise schon gesehen haben, tauchen hier wieder auf.

„Archivierung“ wurde als iterativ gekennzeichnet. Zunächst soll der Name der archivierenden Stelle vermerkt werden. Es folgt ein Feld für Informationen über die vorgesehene Archivierungsdauer. Hier könnte z.B. „Bis 2018“ oder „Langfristig“ stehen. Die Nutzungsrechte der archivierenden Stelle können von der ausschließlichen wissenschaftlichen Auswertung durch den Aufnahmeleiter bis hin zur Veröffentlichung einer Aufnahme im Internet reichen. Sperrvermerke, wie z.B. „Bis 2010 für Externe gesperrt“, können die Nutzungsmöglichkeiten einschränken.

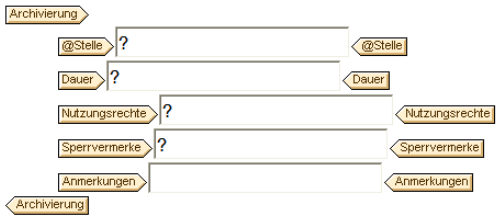

Abb. 104, Korpusbestandteile - Quellaufnahmen - Archivierung

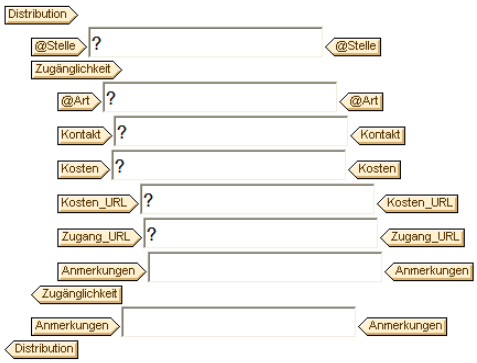

Abb. 105, Korpusbestandteile - Quellaufnahmen - Distribution

„Distribution“ ist ebenfalls iterativ und umfasst neben einem Feld für den Namen der für die Distribution zuständigen Stelle die iterative Komponente „Zugänglichkeit“. In dieser Komponente sollen folgende Angaben verzeichnet werden: Art der Zugänglichkeit, E-Mail-Kontaktadresse, Angaben über die Kosten, ggf. eine URL dieser Angaben sowie ggf. eine URL, die einen direkten Zugang zu den Aufnahmen ermöglicht.

8.3.2. Sprecheignisspezifische Aufnahmen

Neben Quellaufnahmen erfassen wir sprechereignisspezifische Aufnahmen, (SE-Aufnahmen), die in einem bestimmten Verhältnis zu den Quellaufnahmen stehen. Das können Segmente in den Quellaufnahmen bzw. Kopien dieser Segmente sein. Es kommt allerdings auch vor, dass alle SE-Aufnahmen mit den Quellaufnahmen übereinstimmen. Für solche Fälle haben wir die Struktur des iterativen Komplexes SE-Aufnahmen besonders flexibel gestaltet.

Abb. 106, Korpusbestandteile - SE-Aufnahmen (1)

Der Typ der Aufnahmen (Audio, Video und ggf. Tonkopie von Video) sollte im gleichnamigen Feld notiert werden.

8.3.2.1. Basisdaten

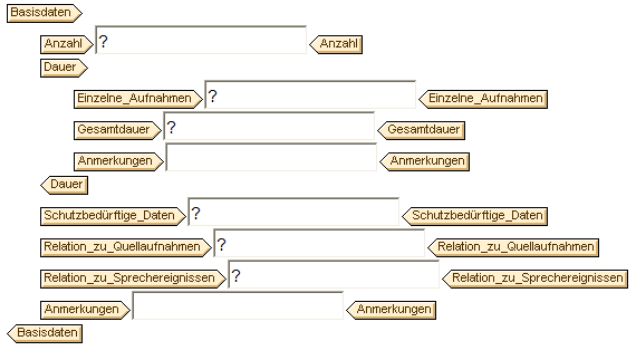

Abb. 107, Korpusbestandteile - SE-Aufnahmen – Basisdaten

Die Anzahl der SE-Aufnahmen des genannten Typs kann man im gleichnamigen Feld notieren. Im nächsten Schritt werden Informationen über die Dauer einzelner Aufnahmen (z.B. „zwischen 4 und 50 Minuten“) und die Gesamtdauer der Aufnahmen erwartet.

Die Felder „Dauer“ und „Schutzbedürftige_Daten“ sind in diesem Komplex fakultativ, d.h. dass sie bei der Erstellung korpusspezifischer Schemata übergangen werden können, wenn Quellaufnahmen dokumentiert wurden und sich die SE-Aufnahmen eines Korpus von den Quellaufnahmen nicht unterscheiden.

Im Feld „Relation_zu_Quellaufnahmen“ sollte vermerkt werden, ob die SE-Aufnahmen mit den Quellaufnahmen übereinstimmen oder ob es sich um Segmente in den Quellaufnahmen handelt. Sprechereignisse können in SE-Aufnahmen vollständig oder unvollständig aufgezeichnet sein. Für solche Angaben gibt es das Feld „Relation_zu_Sprechereignissen“.

8.3.2.2. Transkribierte SE-Aufnahmen

Damit dokumentiert werden kann, wie viele SE-Aufnahmen und welche Arten von Sprechereignissen transkribiert sind, und für eine Information über die Dauer der transkribierten Aufnahmen wurde das Modul „Transkribierte_SE-Aufnahmen“ eingefügt.

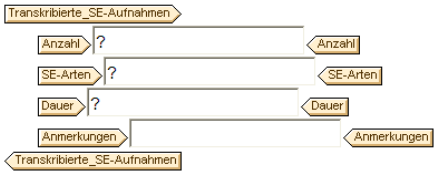

Abb. 108, Korpusbestandteile - SE-Aufnahmen - Transkribierte SE-Aufnahmen

8.3.2.3. Technische Fassungen

Der Komplex „SE-Aufnahmen“ enthält wie der Komplex „Quellaufnahmen“ die fakultativen und iterativen Module „Analoge_Fassungen“ und „Digitale_Fassungen“. Auch diese Teile können übergangen werden, wenn die Quellaufnahmen dokumentiert wurden und sich die SE-Aufnahmen eines Korpus von den Quellaufnahmen nicht unterscheiden.

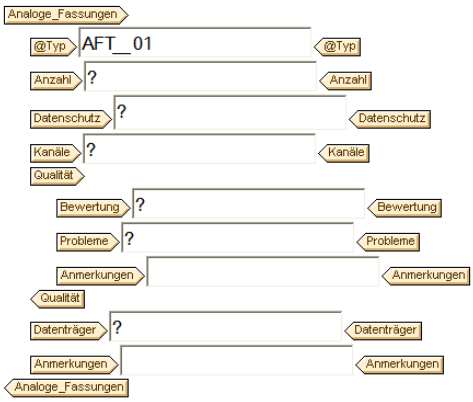

Abb. 109, Korpusbestandteile - SE-Aufnahmen - Analoge Fassungen

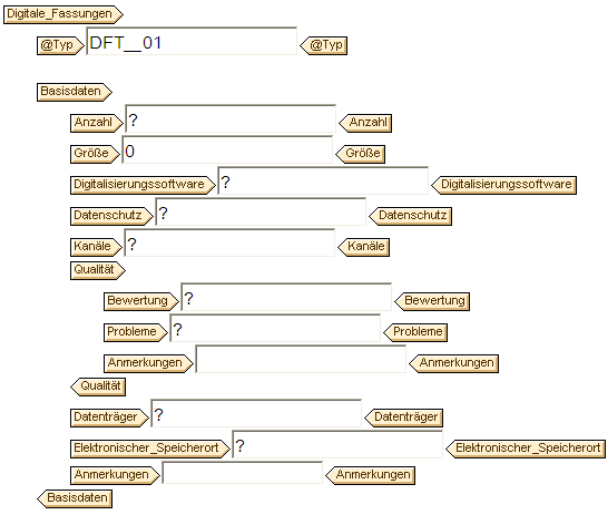

Abb. 110, Korpusbestandteile - SE-Aufnahmen - Digitale Fassungen (1)

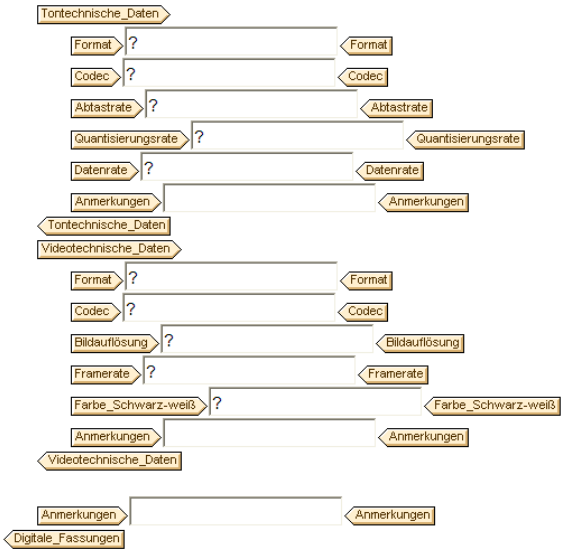

Abb. 111, Korpusbestandteile - SE-Aufnahmen - Digitale Fassungen (2)

Die in den Abb. 109 bis 111 gezeigten Strukturen stimmen mit den in Abschnitt 8.4.1.3. erläuterten überein.

8.3.2.4. Archivierung und Distribution

Die Komponenten „Archivierung“ und „Distribution“ wurden zuletzt in Abschnitt 8.3.1.4. vorgestellt. Wenn die Quellaufnahmen dokumentiert wurden und sich die SE-Aufnahmen eines Korpus von den Quellaufnahmen nicht unterscheiden, können diese Module bei der Beschreibung von SE-Aufnahmen übergangen werden.

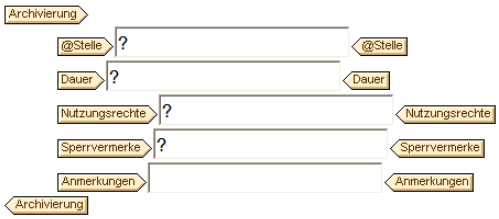

Abb. 112, Korpusbestandteile - SE-Aufnahmen - Archivierung

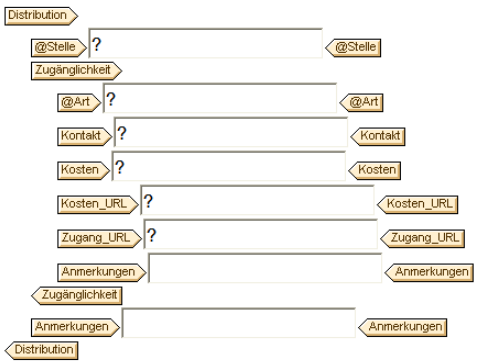

Abb. 113, Korpusbestandteile - SE-Aufnahmen - Distribution

8.3.3. Transkripte

Zu den Korpusbestandteilen zählen auch Transkripte, die allerdings nicht in jedem Korpus enthalten sind, weshalb der Transkriptkomplex im Schema als fakultativ gekennzeichnet wurde.

Abb. 114, Korpusbestandteile - Transkripte (1)

An erster Stelle des Transkriptkomplexes wird die Bezeichnung eines Transkripttypes erwartet. Die Typisierung kann zum einen über die Extensionen (vollständige Transkripte vs. Teiltranskripte) erfolgen, zum anderen über Art und Anzahl der Annotationen. Typenbezeichnungen wie die in Abb. 114 werden auch bei der Dokumentation einzelner Transkripte verwendet.

8.3.3.1. Basisdaten

Im Modul „Basisdaten“ der Transkriptdokumentation kann man die Anzahl der Transkripte des genannten Typs, einen Hinweis auf möglicherweise in den Transkripten enthaltene schutzbedürftige Daten sowie eine Information darüber notieren, ob es sich um vollständige Transkripte oder Teiltranskripte von SE-Aufnahmen handelt.

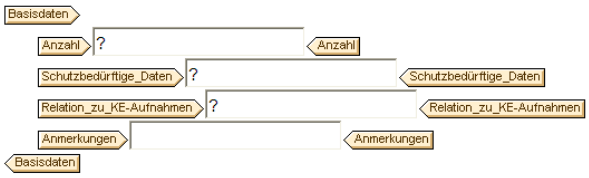

Abb. 115, Korpusbestandteile - Transkripte - Basisdaten

8.3.3.2. Annotationen

In der Beschreibung des Moduls „Annotationen“ greifen wir Erläuterungen aus dem Abschnitt 5.7.2. wieder auf.

Wir verwenden die Bezeichnung „Annotation“ für inhaltlich und formal charakterisierte Ebenen eines Transkripts, wie z.B. Aufzeichnungen des Wortlauts in orthographischer, literarischer oder phonetischer Umschrift, syntaktische Angaben, Notationen suprasegmentaler oder nonverbaler Phänomene, Übersetzung des Wortlautes etc. [14] Da u.U. mehrere Annotationen zu dokumentieren sind, wurde der Komplex im Schema als iterativ gekennzeichnet. [15]

Abb. 116, Korpusbestandteile - Transkripte - Annotationen - Typ

Für jeden ermittelten Annotationstyp wird eine Bezeichnung generiert, die sich zusammensetzt aus dem Kürzel ANT (für „Annotation_Typ“) und einer zweistelligen Nummer. Ein Beispiel finden Sie in Abb. 116. Diese Typenbezeichnungen werden auch in die Dokumentation einzelner Transkripte eingesetzt.

8.3.3.2.1. Basisdaten

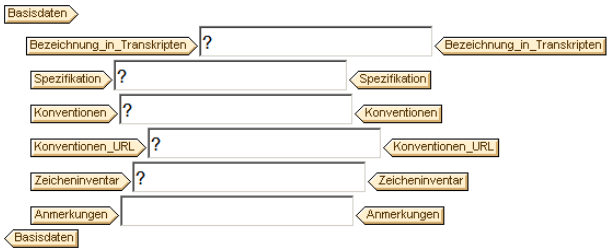

Abb. 117, Korpusbestandteile - Transkripte - Annotationen - Basisdaten

Im ersten Feld der Basisdaten kann man die in Transkripten verwendete Bezeichnung der Annotation notieren. Eine „Spezifikation“ der Annotationen des genannten Typs sollte Angaben über den Gegenstand (z.B. „Wortlaut“), die Umschrift (z.B. „Literarisch“) und die Reichweite (z.B. „Ohne Interviewerbeiträge“) enthalten.

Auf die dem jeweilgen Annotationstyp zugrunde liegenden Konventionen ist im gleichnamigen Feld hinzuweisen. Beispiele für solche Hinweise wären: „Projektspezifisch“, „DIDA, Version vom Januar 2001“, „cGAT“ etc. Unter „Zeicheninventar“ ist das Inventar an Schriftzeichen zu verstehen, das bei der Wiedergabe des Wortlautes verwendet wurde. Das sind i.d.R. standardisierte Inventare wie z.B. der IPA-Zeichensatz oder ein spezifisches Alphabet.

8.3.3.2.2. Erstellung

Auch im iterativen Modul „Erstellung“ der Korpusbeschreibung werden Typen benannt. Für die Typisierung relevant sind die Spezifikationen der Erstellungsprozesse. Abb. 118 enthält ein Beispiel für eine Typenbezeichnung, die auch bei der Dokumentation einzelner Transkripte eingesetzt wird.

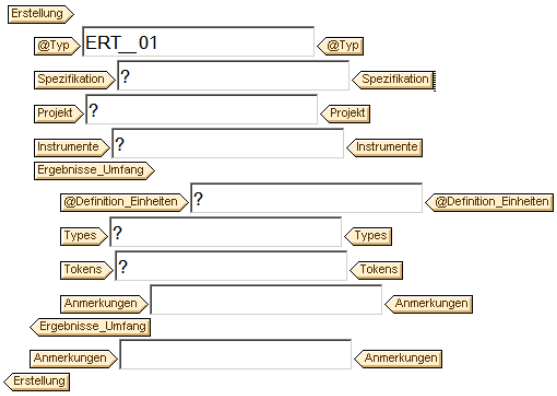

Abb. 118, Korpusbestandteile - Transkripte - Annotationen - Erstellung

Im Feld „Spezifikation“ sollten v.a. der Arbeitsstand (z.B. „Ersterfassung“, „1. Korrektur“, „Endkorrektur“, „Überarbeitung für Publikation xy“) und mögliche besondere Umstände (z.B. „halbautomatisch“) der Erstellung dokumentiert werden. Dann folgen Fragen nach dem für die Erstellung des genannten Typs zuständigen Projekt sowie die bei der Erstellung genutzten Instrumente (Editoren und ggf. Systemumgebung).

Die Information über den Umfang der Ergebnisse eines Erstellungstyps umfasst eine Definition der gezählten Einheiten sowie Felder für Angaben über die Anzahl unterschiedlicher Einheiten (Types) und die Anzahl aller gezählten Einheiten (Tokens). Der Komplex wurde im Schema als iterativ gekennzeichnet.

8.3.3.2.3. Alignment

Wir verwenden die Bezeichnung „Alignment“ in der Dokumentation für die Text-Ton-Synchronisation, also die Koppelung von Aufnahmen und Transkripten auf Phon-, Phonem-, Wort- oder Phrasenbasis, wobei Transkriptsegmenten Zeitmarken zugeordnet werden. Die entsprechende Komponente ist fakultativ und iterativ.

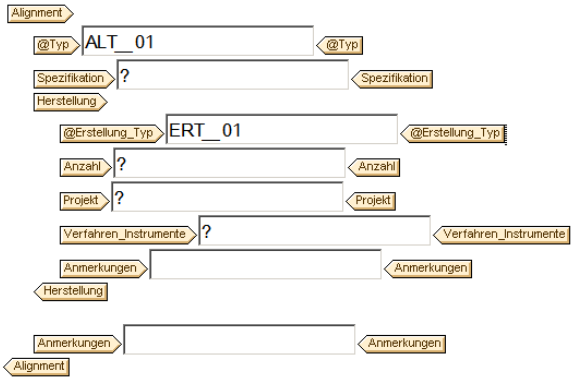

Abb. 119, Korpusbestandteile - Transkripte - Annotationen - Alignment

Auch im Modul „Alignment“ ist eine Typisierung vorgesehen, die sich auf die Spezifikation von Alignmentprozessen stützt. Die hier definierten Typen sollen auch bei der Dokumentation einzelner Transkripte berücksichtigt werden. Ein Beispiel für eine Typenbezeichnung finden Sie in Abb. 119. Im Feld „Spezifikation“ werden Angaben über die für den genannten Typ relevanten Segmente (z.B. „Phonweise“, „Wortweise“) erwartet.

Die Komponente „Herstellung“ ist iterativ. Zunächst wird der Typ der Erstellungen verzeichnet, deren Ergebnisse aligniert wurden. Im Feld „Projekt“ wird der Namen des Projekts, in dem das Alignment vorgenommen wurde, erwartet. Im Feld „Verfahren_Instrumente“ kann man Angaben darüber machen, ob manuell oder automatisch aligniert wurde, auf die genutzte Software hinweisen und ggf. weitere Informationen über die Systemumgebung erfassen.

8.3.3.3. Technische Fassungen

Um die technischen Fassungen von Transkripten dokumentieren zu können, wurden auch in den Transkriptkomplex die Module „Analoge_Fassungen“ und „Digitale_Fassungen“ eingefügt. Beide Abschnitte sind fakultativ und iterativ, wenigstens ein Abschnitt muss bei der Erstellung projektspezifischer Schemata übernommen werden.

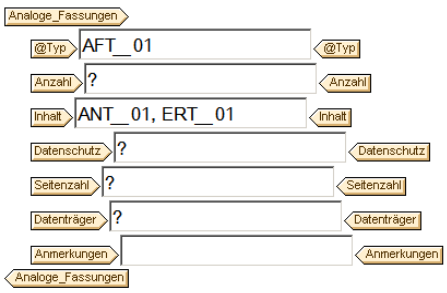

Abb. 120, Korpusbestandteile - Transkripte - Analoge Fassungen

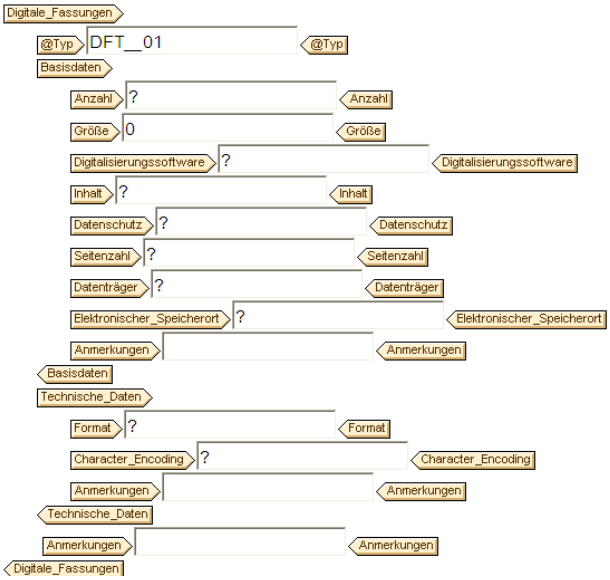

Abb. 121, Korpusbestandteile - Transkripte - Digitale Fassungen

In jedem Abschnitt wird ein Typ abgefragt. Bei analogen Fassungen sind die Werte in den Feldern „Inhalt“ und „Datenschutz“ für eine Typisierung relevant, bei digitalen Fassungen auch Informationen über die technischen Daten (z.B. das Dateiformat). Die Typenbezeichnungen werden auch bei der Dokumentation einzelner Transkripte verwendet. Die Anzahl der Fassungen des genannten Typs kann man im gleichnamigen Feld notieren.

Im Feld „Größe“ kann man den Gesamtumfang der jeweiligen digitalen Fassungen (in Bytes) angeben. Im Feld „Digitalisierungssoftware“ kann man über das Programm, mit dem eine analoge Fassung digitalisiert wurde, informieren. Im iterativen Feld „Inhalt“ sollte man alle Annotationen sowie die Typen der Erstellungen und der Alignments verzeichnen, deren Ergebnisse in den technischen Fassungen des genannten Typs gespeichert sind. „Datenschutz“ meint technische Maßnahmen zum Datenschutz, wie z.B. die Maskierung von Personennamen in Texten. Bei seitenformatierten Texten gibt eine Information über die Seitenzahl einen groben Überblick über den Umfang des Materials. Nach einer Information über den oder die Datenträger wird der elektronische Speicherort der dokumentierten digitalen Fassungen erfasst. An dieser Stelle kann man eine URL oder einen Pfadnamen eintragen. Das erste Feld im Abschnitt „Technische_Daten“ ist für den Namen des Datenformats vorgesehen. „Character_Encoding“ steht für die Zeichencodierung (z.B. ASCII oder UTF-16BE).

8.3.3.4. Archivierung und Distribution

In den Abb. 122 und 123 sehen Sie die Strukturen der schon bekannten Bausteine „Archivierung“ und „Distribution“, die zuletzt in Abschnitt 8.3.1.4. erläutert wurden.

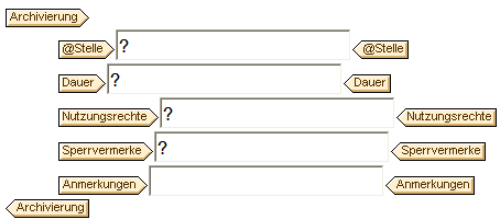

Abb. 122, Korpusbestandteile - Transkripte - Archivierung

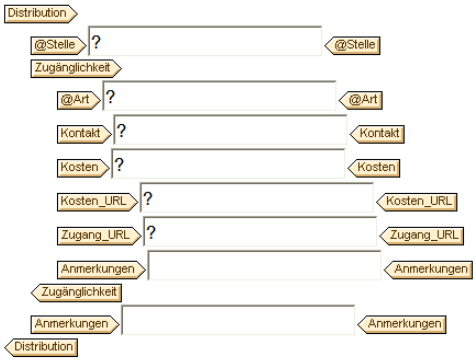

Abb. 123, Korpusbestandteile - Transkripte - Distribution

8.3.4. Zusatzmaterial

Unter „Zusatzmaterial“ verstehen wir Dokumente, die zusätzlich zu Aufnahmen und Transkripten vorhanden sein können. Zusatzmaterialien können auf verschiedenen dokumentarischen Ebenen angesiedelt sein: auf der Ebene der aufgezeichneten Ereignisse (z.B. Skizzen von Sitzordnungen), der Sprechereignisse (z.B. Ablaufprotokolle), der Sprecher (z.B. Sprecherfotos) und auf der Korpusebene (z.B. Transkriptionskonventionen). Der Komplex „Zusatzmaterial“ wurde im Schema als fakultativ und iterativ gekennzeichnet, d.h. dass er bei der Erstellung korpusspezifischer Schemata übergangen und, wenn er gewählt wird, bei der Dateneingabe vervielfältig werden kann.

Abb. 124, Korpusbestandteile - Zusatzmaterial (1)

Im ersten Feld dieses Komplexes sollte die Art des Zusatzmaterials (vgl. o.g. Beispiele) beschrieben werden.

8.3.4.1. Basisdaten

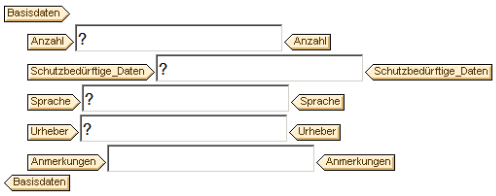

Abb. 125, Korpusbestandteile - Zusatzmaterial - Basisdaten

In der Komponente „Basisdaten“ kann man festhalten, wie viele Zusatzmaterialien der genannten Art vorhanden sind, ob und wenn ja welche schutzbedürftigen Daten diese Dokumente enthalten und welcher Sprache Textdokumente abgefasst wurden. Informationen über den oder die Urheber können im gleichnamigen Feld erfasst werden.

8.3.4.2. Technische Fassungen

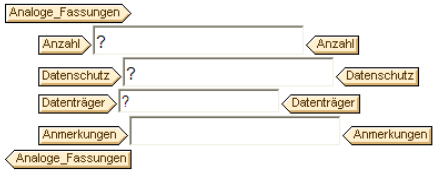

Abb. 126 , Korpusbestandteile - Zusatzmaterial - Analoge Fassungen

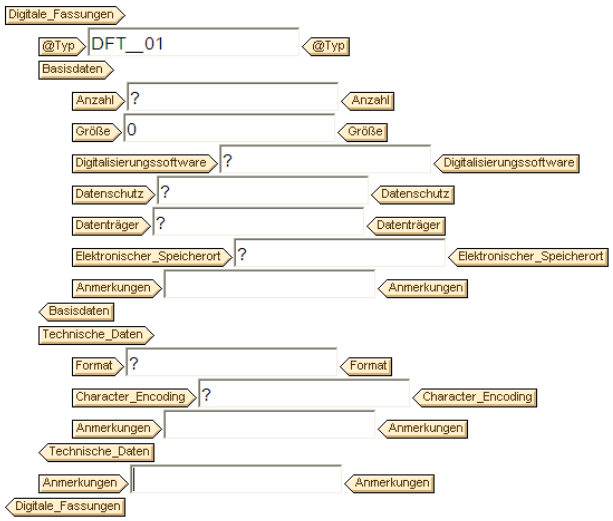

Abb. 127 , Korpusbestandteile - Zusatzmaterial - Digitale Fassungen

Die für die Beschreibung technischer Fassungen von Zusatzmaterial bereitgestellten Module „Analoge_Fassungen“ und „Digitale_Fassungen“ stimmen mit den entsprechenden Modulen im Komplex „Transkripte“ (vgl. 8.3.3.) weitgehend überein.

8.3.4.3. Archivierung und Distribution

Auch mit den Bausteinen „Archivierung“ und „Distribution“ haben wir Sie schon bekannt gemacht, Erläuterungen finden Sie in 8.3.1.4.

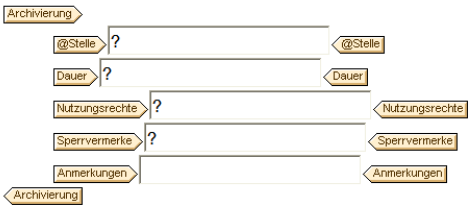

Abb. 128, Korpusbestandteile - Zusatzmaterial - Archivierung

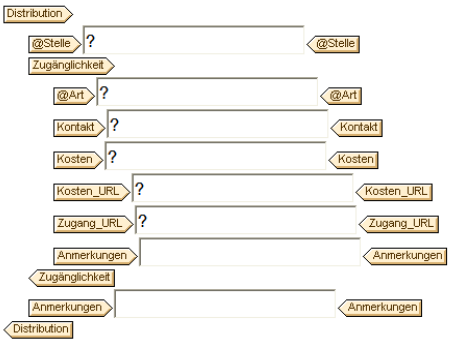

Abb. 129, Korpusbestandteile - Zusatzmaterial - Distribution

8.4. Dokumentationsgeschichte

Informationen über Arbeitsstand und Bearbeiter der Dokumente werden bei der manuellen Dateneingabe automatisch in einer (Oracle-)Datenbank gespeichert, sollten nach unserer Vorstellung jedoch auch in den Dokumenten sichtbar sein. Daher haben wir am Ende aller Schemata den Baustein „Dokumentationsgeschichte“ eingebaut.

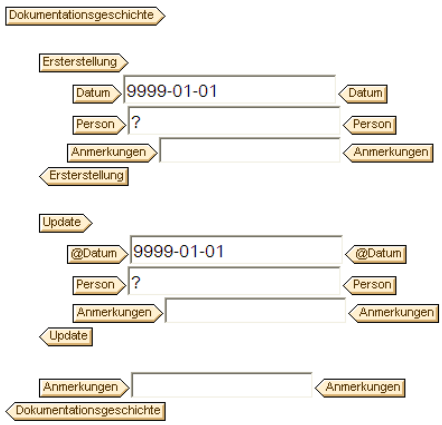

Abb. 130, Dokumentationsgeschichte

Mit dieser letzten Ansicht auf die Struktur des Moduls „Dokumentationsgeschichte“ beenden wir unsere Metadatenbeschreibung und gehen über zu einigen abschließenden Bemerkungen.

9. Abschließende Bemerkungen

Wir haben im vorliegenden Text Grundlagen der Metadatenkomponente der Datenbank für Gesprochenes Deutsch (DGD 2.0) vorgestellt. Im Mittelpunkt der Beschreibung standen vier generische (XML-)Metadaten-Schemata - das Schema für die Dokumentation von Korpusbestandteilen auf Ereignisebene, kurz „Ereignisschema“, das Schema für die Dokumentation ereignis- und sprechereignisübergreifender Sprecherdaten, kurz „Sprecherschema“, das Schema für die Dokumentation von Zusatzmaterial auf Korpusebene sowie ein Schema für überblicksartige Korpusbeschreibungen. In allen Fällen handelt es sich um weitreichende Kategoriensammlungen in flexiblen Strukturen, von denen korpusspezifische Subschemata abgeleitet werden können.

Die auf den oben vorgestellten Schemata basierenden korpusspezifischen Dokumente dienen in erster Linie der projekt- und archivinternen Dokumentation. Für die Außendarstellung von

Korpora, Korpusbestandteilen und Sprechern, über die Externe informiert werden können, werden reduzierte Ansichten entwickelt, d.h. dass nicht alle Daten sichtbar werden..

Auf eine Besonderheit der Entwicklung möchten wir an dieser Stelle noch einmal aufmerksam machen: Wir orientierten uns zunächst u.a. an der ISLE MetaData Initiative (IMDI) [5], die Konventionen für die Veröffentlichung von Metadaten linguistischer Ressourcen vorgelegt hat, sind aber angesichts anderer Aufgaben zu anderen Ergebnissen gekommen als IMDI.

Wir setzten auf vergleichsweise strikte dokumentarische Vorgaben, die wir aus mehreren Gründen für nötig erachten. Die vielfältigen technologischen Möglichkeiten, große Korpora mit einzelnen Bestandteilen in unterschiedlichen Fassungen aufzubauen, spezielle Erfordernisse der Langzeitarchivierung digitaler Daten sowie das Bestreben, solche Bestandteile projektübergreifend nutzbar zu machen, führen zu hohen Anforderungen an die Qualität von Metadaten, mit denen einzelne Projekte nach unserer Erfahrung i.d.R. überfordert sind. Hier besteht ein Regelungsbedarf, dem wir gerecht werden wollen.

Im Rahmen der Begutachtung eines Vortragskonzepts wurden wir gefragt: „Are you suggesting this annotation as a standard?“ Die Antwort lautet: Ja, wir möchten Standards für die Dokumentation von IDS-Korpora der gesprochenen Sprache bereitstellen und hoffen, korpuserstellende Projekte für unser Konzept gewinnen zu können.

10. Anmerkungen

[1] Dublin Core Metadata Initiative. http://dublincore.org/

[2] Simons, Gary & Steven Bird (Hg.) (2006): OLAC Metadata. http://www.language-archives.org/

[3] Text Encoding Initiative. http://www.tei-c.org/

[4] Martínez, José M. (Hg.) (2004): MPEG-7 Overview (version 10). http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

[5] ISLE MetaData Initiative. http://www.mpi.nl/IMDI/

[6] Trippel, Thorsten & Tanja Baumann (2003): Metadaten für Multimodale Korpora: Verwendung im Modelex-Projekt. Technisches Dokument 4, Universität Bielefeld. http://www.spectrum.uni-bielefeld.de/modelex/publication/techdoc/modelex_techrep4/

Die Originalformatierung der zitierten Textpassage wurde aus Platzgründen nicht übernommen.

[7] IMDI, Part 1, Metadata Elements for Session Description, Version 3.0.4, October 2003, S. 10. http://www.mpi.nl/IMDI/

[8] Schiel, Florian & Christoph Draxler (2004): The production of speech corpora. Version 2.5. http://www.phonetik.uni-muenchen.de/forschung/BITS/TP1/Cookbook/

[9] Gesamtkatalog der Tonaufnahmen des Deutschen Spracharchivs. Erarbeitet von Mitarbeiterinnen und Mitarbeitern des Instituts für Deutsche Sprache. Phonai Bde. 38 u. 39 - Tübingen: Niemeyer, 1992.

[10] Schreibkonventionen findet man im „Regelwerk Mediendokumentation“. http://rmd.dra.de/arc/php/main.php (Mitteilung von Ulf-Michael Stift)

[11] Wir nennen hier lediglich zur Veranschaulichung einige Produkte: Adobe Audition, Audacity, Audiograbber, Digidesign ProTools, No23 Recorder, Steinberg Wavelab.

[12] Diese Informationen bekamen wir von Jürgen Immerz.

[13] http://agd.ids-mannheim.de/html/korpora/pdf/isdok.pdf

[14] Das bedeutet, dass hier unter “Annotation” nicht etwa die einzelne, segmentbezogene und auf einer semantisch definierten und einem Sprecher zugeordneten Annotationsspur eingetragene „Annotation“ verstanden wird, wie der Sprachgebrauch in ELAN wäre. (Mitteilung von Wilfried Schütte)

[15] Verschiedenen Annotationen können vermischt sein (z.B. Wortlaut in literarischer Umschrift plus Intonationsnotation). In solchen Fällen empfiehlt es sich, in der Dokumentation mit einem Annotationskomplex zu arbeiten, in dem alle Angaben zusammengefasst werden. Wenn verschiedenen Annotationen getrennt gehalten sind - z.B. a) Wortlaut in literarischer Umschrift plus Intonationsnotation, b) morphosyntaktische Annotation, c) Notation nonverbaler Phänomene - sollten für a), b) und c) verschiedene Annotationskomplexe angelegt werden.

[16] Caren Brinckmann, Stefan Kleiner, Ralf Knöbl and Nina Berend (2008): German Today: an areally extensive corpus of spoken Standard German. In: Proceedings 6th International Conference on Language Resources and Evaluation (LREC 2008), Marrakesch, Marokko.

http://www.lrec-conf.org/proceedings/lrec2008/summaries/806.html

[17] Florian Schiel, Angela Baumann, Christoph Draxler, Tania Ellbogen, Phil Hoole, Alexander Steffen: The Validation of Speech Corpora. Version 1.11 : June 3, 2004. http://www.phonetik.uni-muenchen.de/forschung/BITS/TP2/Cookbook/Tp2.html

[18] http://www.ids-mannheim.de/oea/forsch/