DGD 2.0: A Web-based Navigation Platform for the Visualization, Presentation and Retrieval of German Speech Corpora

1. Introduction

1.1 The Collection of German Speech Corpora at the IDS

The "Institut für Deutsche Sprache" (IDS) is hosting a wide range of historical and contemporary German speech corpora. Many of the historical corpora, especially the dialectological corpora like for example the Zwirner Corpus[1] can be accessed online via the "Database for Spoken German" (DGD)[2]. Currently we are developing a new, generic speech corpus management system at the IDS where the normalized integration of historical and more recent speech corpora under sustainability aspects and an object-oriented user interface[3] for corpus exploration and querying are monitored as main objectives. The XML schema-based standardization on meta documentation and transcript data level allows the implementation of exact mapping mechanisms for the import and export of existing and future speech corpora. Furthermore, the specific characteristics of individual speech corpora are preserved.

1.2 The Standardization Approach for cross-Corpus Information Management

The theoretical approach of the XML schema based standardization of speech corpus meta data components has been previously described in detail using the example of the corpus instance "German Today"[4].

All speech corpus systems manage meta-information regarding their media source signals. However the information structures of the data components used in different speech corpora may vary considerably regarding the linguistic research questions that are investigated by their creators. Such differences between speech corpora can originate for example from the represented genres, from the degree of content restriction, from the physical data structure or from the research field focussed on (i.e. natural vs. elicited speech)[5].

This article describes our Web-based speech corpus navigation platform that focuses on an abstract standardization concept – matching large speech corpus collections rather than creating particular solutions for data sets of single speech corpus projects. The cross-corpus perspective leads to the definition of a generic, system-wide data model, allowing a smooth corpus data integration without information loss. The components of this data model are hierarchically interlinked with each other as shown in figure 1: the structured XML documentation instances on corpus, event and speaker level, the unstructured, semi-structured or time aligned transcripts, the media sources and perhaps additional unstructured secondary documents are interrelated by system-wide, unique identifiers. The output quality of cross-corpus information processing[6] may vary as it is

[1] The Zwirner Corpus was initiated in 1955.

[2] URL: http://dsav-wiss.ids-mannheim.de/

[3] Cf. Nielsen (1994), p. 58 f.

[4] Cf. Gasch (2008), chapter 1.2 „Generic and project-specific XML Schema design", p. 23 f.

[5] Cf. Wichmann (2008), p. 190 ff.

[6] like for example the information retrieval in cross-corpus transcript collections

always directly prescribed by the corpus data component with the lowest structure or data quality.

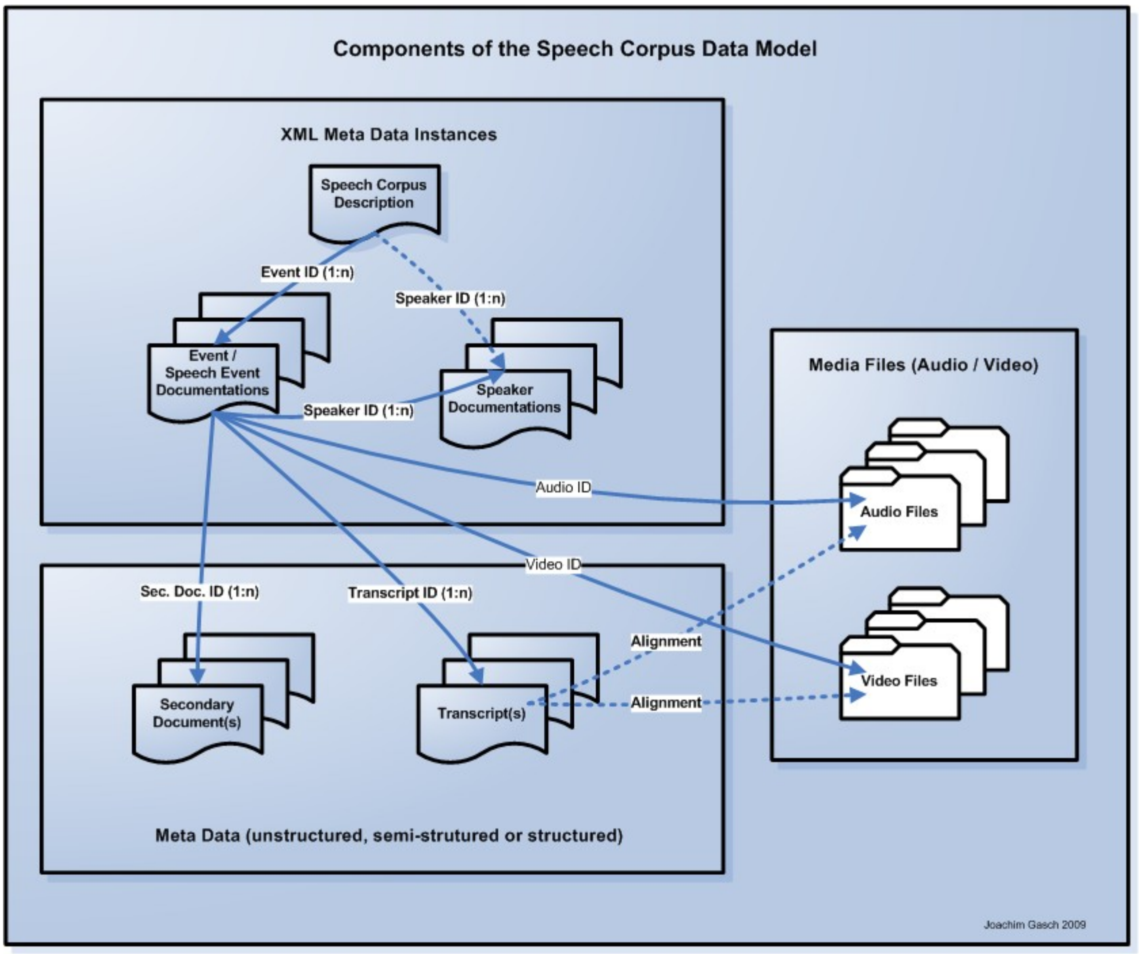


Figure 1: Interlinked components of the normalized speech corpus data model

2. The Online Navigation Platform

2.1 The Navigation Interface – Design Principals

Speech corpus navigation interfaces give access to extensive corpus related meta data and transcripts. Nielsen [1994][7] explains that the structure of graphical interfaces can be guided by either function- or object-oriented design principles. With function-based interfaces, the interface structure is defined by the command set to be implemented which often leads to command-line overloaded interfaces. However the structure of object-oriented interfaces is based on the information units of the documents to be managed by the system. Our speech corpus interface has to provide adaptive views of speech corpus data components. Therefore the document-centric, object-oriented interaction paradigm was chosen to design the application menu and the partitioning of the user's screen.

The flat structured navigation menu represents the different components of the speech corpus data model (as described in chapter 1.2). It is implemented at the top of the screen with a fixed position in order to permanently provide a system-wide,

[7] Cf. Nielsen (1994), p. 64 f.

homogeneous access to application components. The symbols ► and ▼ are indicating whether menu entry points are flat or hierarchically subdivided (cf. figure 2).

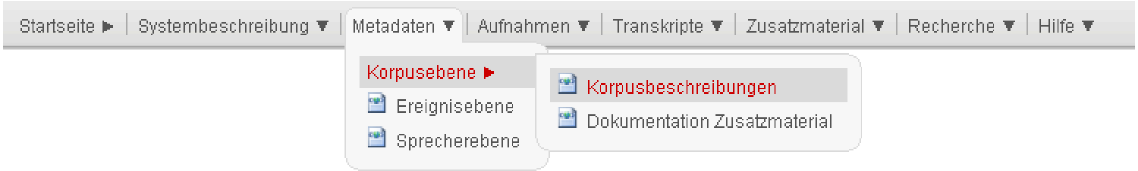


Figure 2: The navigation menu of the user interface

For a more intuitive user orientation, the specific types of data components like XML meta-documentations, audio files, transcripts or secondary documents are marked by corresponding classified icons:

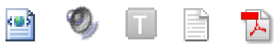
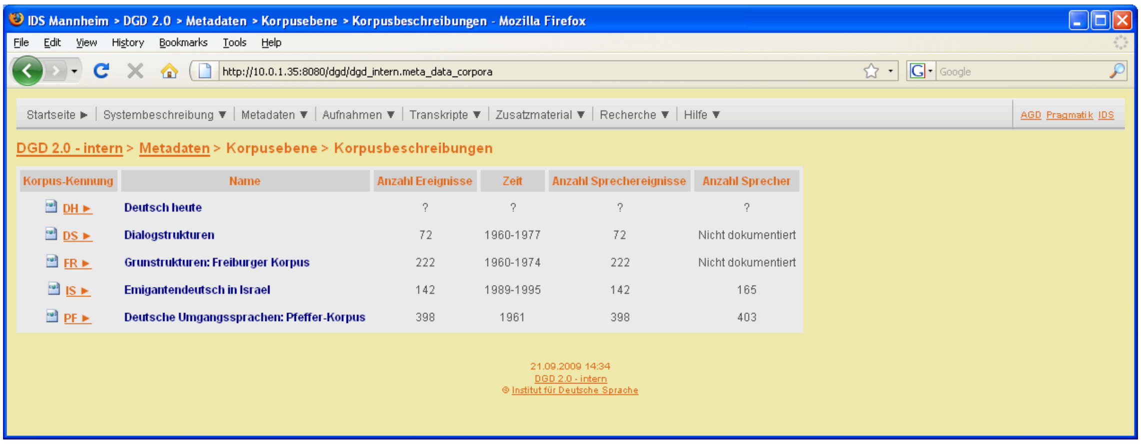


Figure 3: The classified icons for specific types of data components

Additionally, "bread crumb" navigation helps the user at any time to identify his current position in the navigation tree and also to navigate back with one mouse click to the previous application level or to the system start page. Figure 4 shows the current location in the navigation tree ("DGD 2.0 intern > Metadaten > Korpusebene > Korpusbeschreibungen") including the two links ("Metadaten" and "DGD 2.0 intern") for back navigation to the second or the first menu level. A large working area is implemented in the centre of the screen to respond to the need to display large meta-documentations and transcripts.



IDS Mannheim > DGD 2.0 > Metadaten > Korpusebene > Korpusbeschreibungen - Mozilla Firefox

http://10.0.1.35:8080/dgd/dgd_intern.meta_data_corpora

DGD 2.0 - intern > Metadaten > Korpusebene > Korpusbeschreibungen

Korpus-Kennung	Name	Anzahl Ereignisse	Zeit	Anzahl Sprechereignisse	Anzahl Sprecher
DH ►	Deutsch heute	?	?	?	?
DS ►	Dialogstrukturen	72	1960-1977	72	Nicht dokumentiert
FR ►	Grunstrukturen: Freiburger Korpus	222	1960-1974	222	Nicht dokumentiert
IS ►	Emigantendeutsch in Israel	142	1989-1995	142	165
PF ►	Deutsche Umgangssprachen: Pfeffer-Korpus	398	1961	398	403

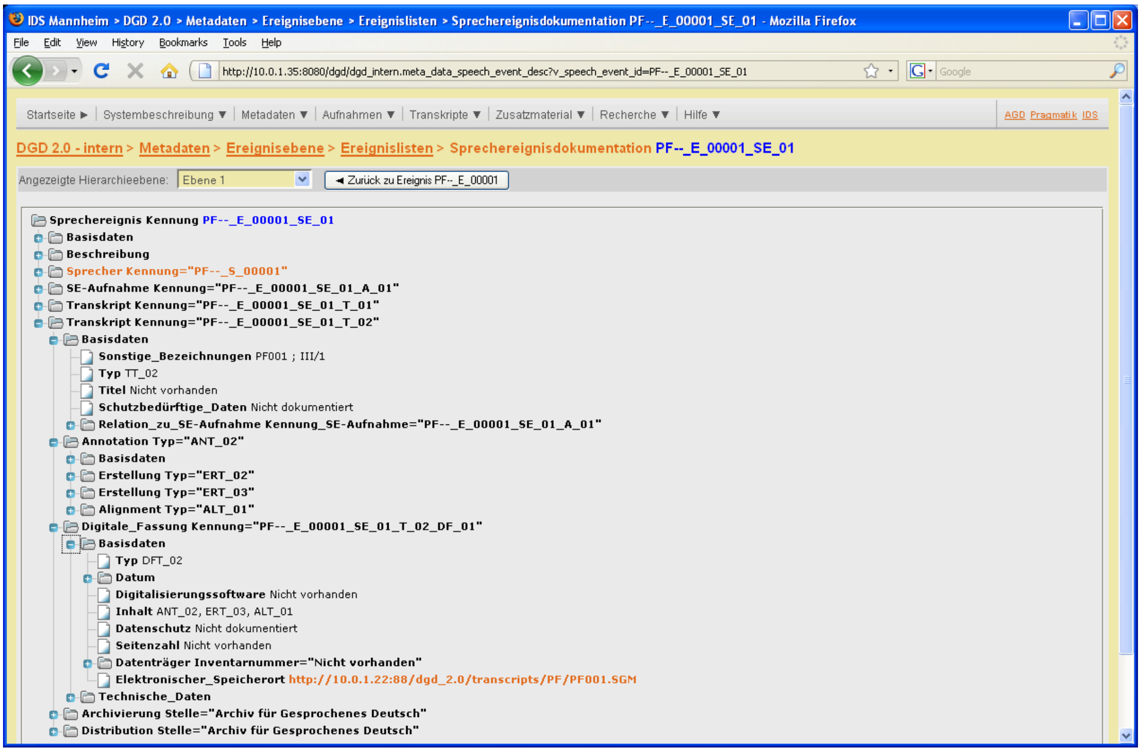
21.09.2009 14:34
DGD 2.0 - intern
© Institut für Deutsche Sprache

Figure 4: "Bread crumb" navigation and working area

2.2 The Visualization and Presentation of Speech Corpus Content

2.2.1 Generic Visualization of the XML Meta-Information of Speech Corpora

All speech corpus related XML meta-documentations are stored natively and valid against the corresponding XML catalogue schema in the system's XML database. To avoid corpus specific instance visualizations, a generic XML document rendering module has been implemented. It provides XML tree visualization, expandable / collapsible document nodes and a node level selection functionality. Hyperlinks to external resources are directly accessible from the XML visualization (cf. figure 5).


IDS Mannheim > DGD 2.0 > Metadaten > Ereignisebene > Ereignislisten > Sprechereignisdokumentation PF--_E_00001_SE_01 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://10.0.1.35:8080/dgd/dgd_intern.meta_data_speech_event_desc?v_speech_event_id=PF--_E_00001_SE_01
Google
Startseite ▶ | Systembeschreibung ▼ | Metadaten ▼ | Aufnahmen ▼ | Transkripte ▼ | Zusatzmaterial ▼ | Recherche ▼ | Hilfe ▼
AGD Pragmatik IDS
DGD 2.0 - intern > Metadaten > Ereignisebene > Ereignislisten > Sprechereignisdokumentation PF--_E_00001_SE_01
◄ Zurück zu Ereignis PF--_E_00001

Angezeigte Hierarchieebene: Ebene 1

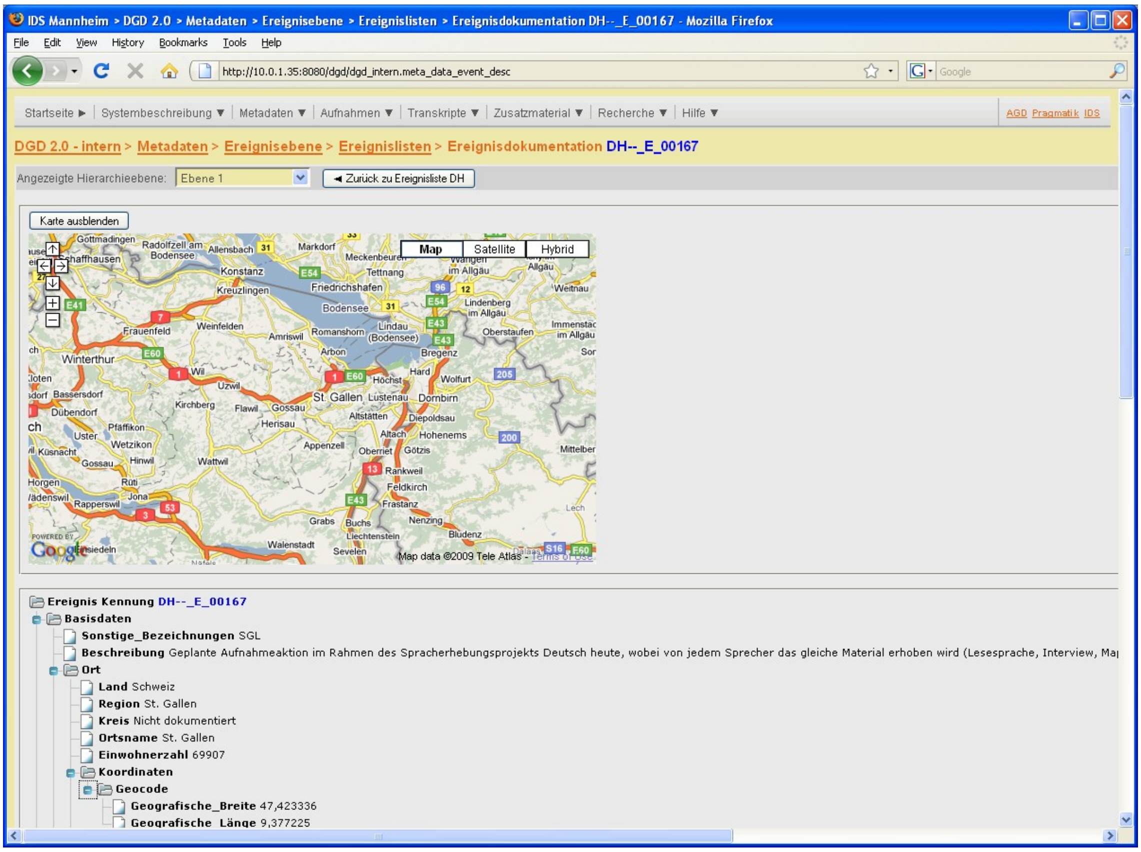
```
Sprechereignis Kennung PF--_E_00001_SE_01
  Basisdaten
  Beschreibung
  Sprecher Kennung="PF--_S_00001"
  SE-Aufnahme Kennung="PF--_E_00001_SE_01_A_01"
  Transkript Kennung="PF--_E_00001_SE_01_T_01"
  Transkript Kennung="PF--_E_00001_SE_01_T_02"
    Basisdaten
      Sonstige_Bezeichnungen PF001 ; III/1
      Typ TT_02
      Titel Nicht vorhanden
      Schutzbedürftige_Daten Nicht dokumentiert
      Relation_zu_SE-Aufnahme Kennung_SE-Aufnahme="PF--_E_00001_SE_01_A_01"
    Annotation Typ="ANT_02"
      Basisdaten
      Erstellung Typ="ERT_02"
      Erstellung Typ="ERT_03"
      Alignment Typ="ALT_01"
    Digitale_Fassung Kennung="PF--_E_00001_SE_01_T_02_DF_01"
      Basisdaten
        Typ DFT_02
        Datum
        Digitalisierungssoftware Nicht vorhanden
        Inhalt ANT_02, ERT_03, ALT_01
        Datenschutz Nicht dokumentiert
        Seitenzahl Nicht vorhanden
        Datenträger Inventarnummer="Nicht vorhanden"
        Elektronischer_Speicherort http://10.0.1.22:88/dgd_2.0/transcripts/PF/PF001.SGM
      Technische_Daten
    Archivierung Stelle="Archiv für Gesprochenes Deutsch"
    Distribution Stelle="Archiv für Gesprochenes Deutsch"
```

Figure 5: Generic XML document rendering

The cross-corpus display method of corpus-, event- and speaker documentation offers an ergonomic navigation experience especially for large data-centric XML documentation instances.

Speech corpus projects also may document the geographic coordinates of their event locations like for example in the case of the project "German Today"[8]. A geographic map can be displayed on demand for each event. Figure 6 shows the geographic map for the event DH--_E_00167 (with geographic latitude 47.423336 and longitude 9.377225) which took place in St. Gallen (Switzerland).

[8] Cf.: http://www.ids-mannheim.de/prag/AusVar/Deutsch_heute/



DGD 2.0 - intern > Metadaten > Ereignisebene > Ereignislisten > Ereignisdokumentation DH--_E_00167

Angezeigte Hierarchieebene: Ebene 1 ◄ Zurück zu Ereignisliste DH

Ereignis Kennung DH--_E_00167
- **Basisdaten**
  - **Sonstige_Bezeichnungen** SGL
  - **Beschreibung** Geplante Aufnahmeaktion im Rahmen des Spracherhebungsprojekts Deutsch heute, wobei von jedem Sprecher das gleiche Material erhoben wird (Lesesprache, Interview, Ma
  - **Ort**
    - **Land** Schweiz
    - **Region** St. Gallen
    - **Kreis** Nicht dokumentiert
    - **Ortsname** St. Gallen
    - **Einwohnerzahl** 69907
    - **Koordinaten**
      - **Geocode**
        - **Geografische_Breite** 47,423336
        - **Geografische_Länge** 9,377225

Figure 6: Geographic map (based on documented geocodes showing the location of an event)

2.2.2 Transcript Visualization and Presentation

The more speech corpora (and especially historical corpora) belong to the corpus collection, the more a common concept of "transcript" becomes fuzzy – in the sense of the annotated phenomena and first of all in the sense of the data formats used.

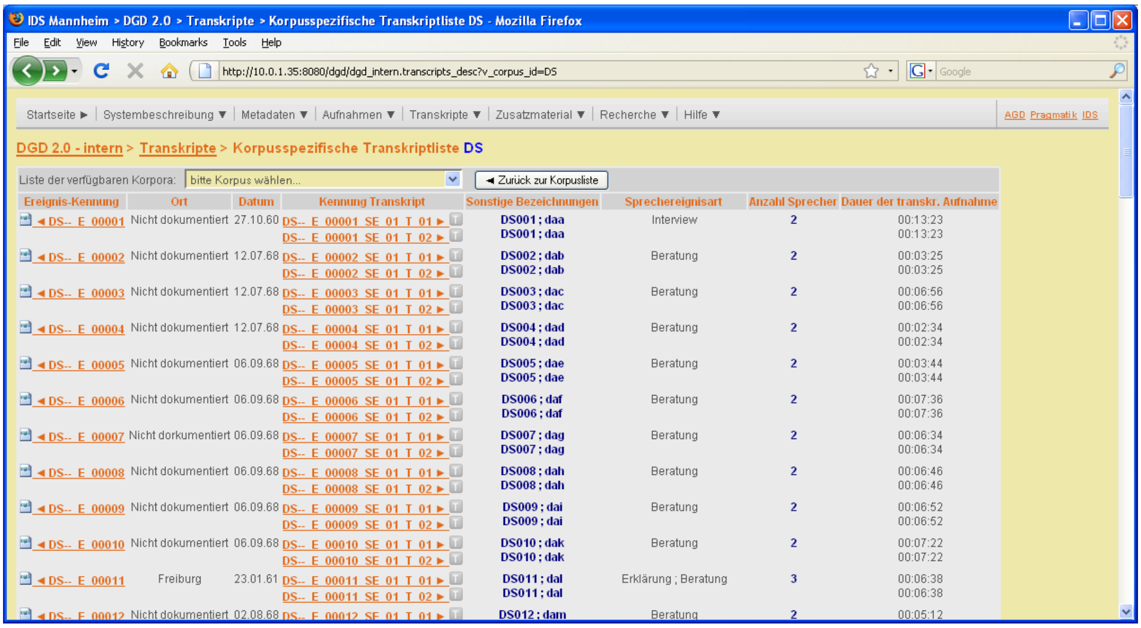
Recent speech corpus projects mostly use one of the annotation tools available nowadays[9]. These tools clearly separate annotation content from its graphical representation and most of them also include time alignment. However, the transcripts of historical corpora consist of project-specific data formats, often exclusively oriented on the graphical layout of annotation content without a time alignment of the annotated information. Therefore our speech corpus management system has to operate with unstructured, purely layout-oriented data formats[10] and with transcript editor specific document structures at the same time. At this point, no generic visualization of transcript information is possible. Concepts here mainly depend on the structuring degree of the transcripts.

If we look at the transcript-related part of the meta-documentation, a generic visualization of information is possible. Corpus-specific transcript lists are accessible via the main menu entry point "Transkripte". Figure 7 shows the corpus-specific transcript access list for the speech corpus DS[11].

[9] A list (in german language) of currently available annotation tools can be found at the following URL: http://prowiki.ids-mannheim.de/bin/view/GAIS/TranskriptionEditoren

[10] like for example text, html or proprietary word processor formats

[11] DS: Dialogstrukturen

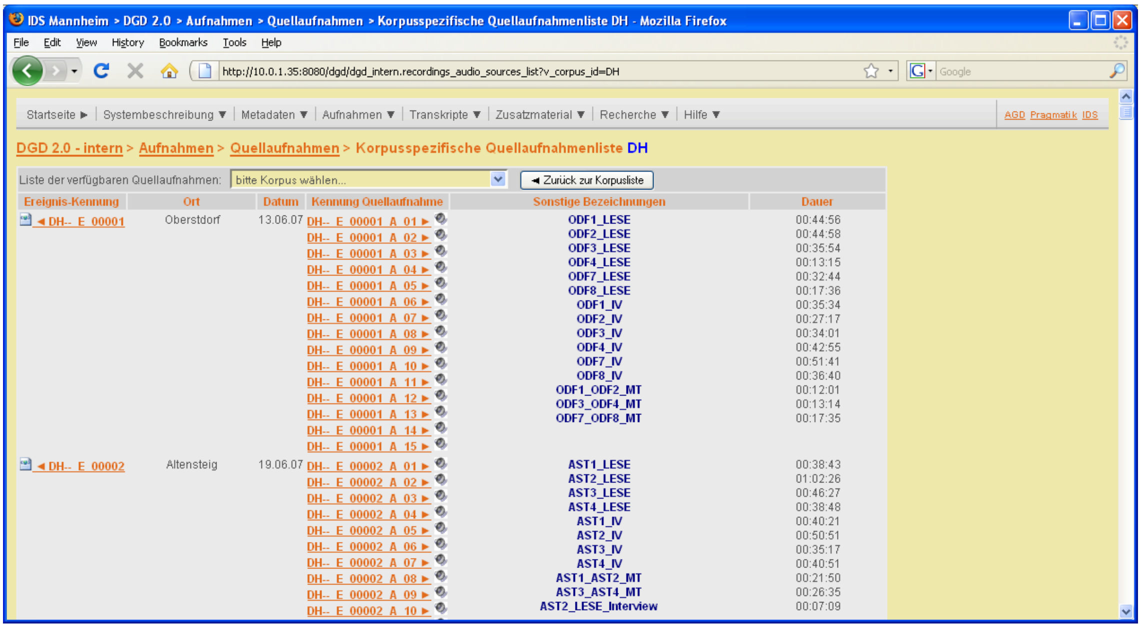

IDS Mannheim > DGD 2.0 > Transkripte > Korpusspezifische Transkriptliste DS - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://10.0.1.35:8080/dgd/dgd_intern.transcripts_desc?v_corpus_id=DS
Google
Startseite ► | Systembeschreibung ▼ | Metadaten ▼ | Aufnahmen ▼ | Transkripte ▼ | Zusatzmaterial ▼ | Recherche ▼ | Hilfe ▼
AGD Pragmatik IDS
DGD 2.0 - intern > Transkripte > Korpusspezifische Transkriptliste DS
Liste der verfügbaren Korpora: bitte Korpus wählen...
◄ Zurück zur Korpusliste

Ereignis-Kennung	Ort	Datum	Kennung Transkript	Sonstige Bezeichnungen	Sprechereignisart	Anzahl Sprecher	Dauer der transkr. Aufnahme
◄ DS--_E_00001	Nicht dokumentiert	27.10.60	DS--_E_00001_SE_01_T_01 ► DS--_E_00001_SE_01_T_02 ►	DS001 ; daa DS001 ; daa	Interview	2	00:13:23 00:13:23
◄ DS--_E_00002	Nicht dokumentiert	12.07.68	DS--_E_00002_SE_01_T_01 ► DS--_E_00002_SE_01_T_02 ►	DS002 ; dab DS002 ; dab	Beratung	2	00:03:25 00:03:25
◄ DS--_E_00003	Nicht dokumentiert	12.07.68	DS--_E_00003_SE_01_T_01 ► DS--_E_00003_SE_01_T_02 ►	DS003 ; dac DS003 ; dac	Beratung	2	00:06:56 00:06:56
◄ DS--_E_00004	Nicht dokumentiert	12.07.68	DS--_E_00004_SE_01_T_01 ► DS--_E_00004_SE_01_T_02 ►	DS004 ; dad DS004 ; dad	Beratung	2	00:02:34 00:02:34
◄ DS--_E_00005	Nicht dokumentiert	06.09.68	DS--_E_00005_SE_01_T_01 ► DS--_E_00005_SE_01_T_02 ►	DS005 ; dae DS005 ; dae	Beratung	2	00:03:44 00:03:44
◄ DS--_E_00006	Nicht dokumentiert	06.09.68	DS--_E_00006_SE_01_T_01 ► DS--_E_00006_SE_01_T_02 ►	DS006 ; daf DS006 ; daf	Beratung	2	00:07:36 00:07:36
◄ DS--_E_00007	Nicht dorkumentiert	06.09.68	DS--_E_00007_SE_01_T_01 ► DS--_E_00007_SE_01_T_02 ►	DS007 ; dag DS007 ; dag	Beratung	2	00:06:34 00:06:34
◄ DS--_E_00008	Nicht dokumentiert	06.09.68	DS--_E_00008_SE_01_T_01 ► DS--_E_00008_SE_01_T_02 ►	DS008 ; dah DS008 ; dah	Beratung	2	00:06:46 00:06:46
◄ DS--_E_00009	Nicht dokumentiert	06.09.68	DS--_E_00009_SE_01_T_01 ► DS--_E_00009_SE_01_T_02 ►	DS009 ; dai DS009 ; dai	Beratung	2	00:06:52 00:06:52
◄ DS--_E_00010	Nicht dokumentiert	06.09.68	DS--_E_00010_SE_01_T_01 ► DS--_E_00010_SE_01_T_02 ►	DS010 ; dak DS010 ; dak	Beratung	2	00:07:22 00:07:22
◄ DS--_E_00011	Freiburg	23.01.61	DS--_E_00011_SE_01_T_01 ► DS--_E_00011_SE_01_T_02 ►	DS011 ; dal DS011 ; dal	Erklärung ; Beratung	3	00:06:38 00:06:38
◄ DS--_E_00012	Nicht dokumentiert	02.08.68	DS--_E_00012_SE_01_T_01 ►	DS012 ; dam	Beratung	2	00:05:12

Figure 7: Corpus-specific transcript list for the speech corpus DS

2.2.3 Media Presentation

Speech corpora may include different types of interdependent media files. First, each corpus event is related to one or more source files. This is the raw material of the recorded event, originating for example directly from an audio recording device. As an event may be composed of several speech events, the original raw material can be further segmented into speech event specific recordings. All relevant information regarding the different media file types is maintained in the meta-documentation of the corresponding event and can be accessed via the list of the main menu entry point "Aufnahmen" as shown in figure 8.


IDS Mannheim > DGD 2.0 > Aufnahmen > Quellaufnahmen > Korpusspezifische Quellaufnahmenliste DH - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://10.0.1.35:8080/dgd/dgd_intern.recordings_audio_sources_list?v_corpus_id=DH
Google
Startseite ► | Systembeschreibung ▼ | Metadaten ▼ | Aufnahmen ▼ | Transkripte ▼ | Zusatzmaterial ▼ | Recherche ▼ | Hilfe ▼
AGD Pragmatik IDS
DGD 2.0 - intern > Aufnahmen > Quellaufnahmen > Korpusspezifische Quellaufnahmenliste DH
Liste der verfügbaren Quellaufnahmen: bitte Korpus wählen...
◄ Zurück zur Korpusliste

Ereignis-Kennung	Ort	Datum	Kennung Quellaufnahme	Sonstige Bezeichnungen	Dauer
◄ DH--_E_00001	Oberstdorf	13.06.07	DH--_E_00001_A_01 ►	ODF1_LESE	00:44:56
			DH--_E_00001_A_02 ►	ODF2_LESE	00:44:58
			DH--_E_00001_A_03 ►	ODF3_LESE	00:35:54
			DH--_E_00001_A_04 ►	ODF4_LESE	00:13:15
			DH--_E_00001_A_05 ►	ODF7_LESE	00:32:44
			DH--_E_00001_A_06 ►	ODF8_LESE	00:17:36
			DH--_E_00001_A_07 ►	ODF1_IV	00:35:34
			DH--_E_00001_A_08 ►	ODF2_IV	00:27:17
			DH--_E_00001_A_09 ►	ODF3_IV	00:34:01
			DH--_E_00001_A_10 ►	ODF4_IV	00:42:55
			DH--_E_00001_A_11 ►	ODF7_IV	00:51:41
			DH--_E_00001_A_12 ►	ODF8_IV	00:36:40
			DH--_E_00001_A_13 ►	ODF1_ODF2_MT	00:12:01
			DH--_E_00001_A_14 ►	ODF3_ODF4_MT	00:13:14
			DH--_E_00001_A_15 ►	ODF7_ODF8_MT	00:17:35
◄ DH--_E_00002	Altensteig	19.06.07	DH--_E_00002_A_01 ►	AST1_LESE	00:38:43
			DH--_E_00002_A_02 ►	AST2_LESE	01:02:26
			DH--_E_00002_A_03 ►	AST3_LESE	00:46:27
			DH--_E_00002_A_04 ►	AST4_LESE	00:38:48
			DH--_E_00002_A_05 ►	AST1_IV	00:40:21
			DH--_E_00002_A_06 ►	AST2_IV	00:50:51
			DH--_E_00002_A_07 ►	AST3_IV	00:35:17
			DH--_E_00002_A_08 ►	AST4_IV	00:40:51
			DH--_E_00002_A_09 ►	AST1_AST2_MT	00:21:50
			DH--_E_00002_A_10 ►	AST3_AST4_MT	00:26:35
				AST2_LESE_Interview	00:07:09

Figure 8: Corpus-specific list of source recordings for the speech corpus DH

3. Retrieval Strategies for unstructured and structured Speech Corpus Data Components

Media file content can not be located without descriptive meta-information[12]. Current speech corpus retrieval systems gather information about media source files by exploring the corresponding meta data and possibly related transcript data. As described in the previous chapter, the transcript data components of speech corpus collections may spread regarding their structuring degree. Appropriate retrieval strategies mainly depend on this degree. For structured transcripts the distinction between a document-centric and a data-centric organization of the annotation layers is also of interest[13]. Document-centric (single-layer) transcripts can be seen as semi-structured transcripts, enabling only full-text search results that are enhanced by additional time alignment information. This also means that a complex layer-aware query processing against transcripts with one specific multi-layer structure will have to be downgraded to a simple full-text search as soon as different (document-centric) structures from other corpora join the same query. In a worst case scenario, the same partial incompatibility effect can even occur with transcripts of one single corpus that were created with a current multi-layer transcript editor. Different reasons may have contributed to inhomogeneous transcript data:

- Signal segmentation without precise segmentation guidelines (i.e. phones, words, phrases or turns)
- No or not sufficient naming conventions applied for the different transcript layer descriptors (i.e. no unique descriptor used for orthographic transcription layer)
- No exact semantic layer definition available or semantic mix-up of layer content (i.e. mix-up of orthographic and phonetic markup in one single layer)
- No exact syntactic definition of layer content available or syntactic mix-up of layer content (i.e. mix-up of punctuation- or capitalization conventions in the orthographic layer)
- Violation of cross-layer time relations (i.e. caused by interval changes that were made with multi-layer transcript editors without layer inheritance control)

All these reasons may cause structural incompatibilities with the result that full-text search is the lowest common denominator regarding feasible retrieval strategies. In an online environment, the determination of the appropriate retrieval strategy has to be an automated process depending on the user's current corpus selection.

3.1 The Full-Text Search Module

The implementation of full-text search functionality does not require structured data. But at the same time structured data can also be included during the search. For semi- or unstructured transcript data only a full-text search interface can be provided. As we already use the Oracle XML Database for our XML meta-documentation storage needs, we can also benefit from the rich full-text search features provided by Oracle Text[14] for

[12] Cf. U.S. General Services Administration (2005), p. 9

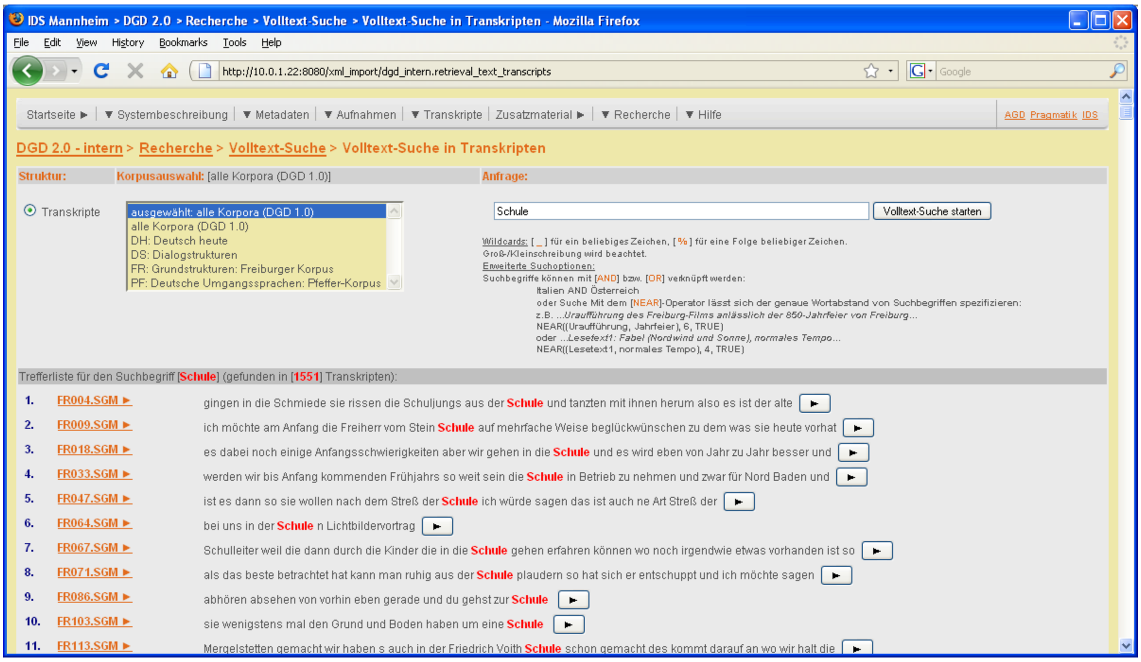
[13] Cf. Trippel (2004), p. 2

[14] Cf. Schneider (2009), p. 144-147, (in german language)

the full-text search in transcripts. Here some examples of the full-text query features provided:

- The simple and multiple wildcard characters "_" and "**%**":
  _ind matches i.e. "Kind" and "Wind"
  %wind matches i.e. "Nordwind" or Südwind"
- the operators **AND** and **OR** build logical relations between search terms:
  Nordwind AND Südwind matches only documents with occurrences of both terms
- the **NOT** operator excludes a specific search term:
  Nordwind NOT Südwind matches only documents containing "Nordwind" but not containing "Südwind"
- the **NEAR** operator finds documents depending on the word distance of search terms:
  NEAR((Schule, Kirche, 4, true) matches documents where both search terms occur with a (maximum) word distance of 4 words.

A working application example is shown in figure 9 (the audio segments corresponding to the query hits are directly accessible by pressing the corresponding button).


IDS Mannheim > DGD 2.0 > Recherche > Volltext-Suche > Volltext-Suche in Transkripten - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://10.0.1.22:8080/xml_import/dgd_intern.retrieval_text_transcripts
Google
Startseite ▶ | ▼ Systembeschreibung | ▼ Metadaten | ▼ Aufnahmen | ▼ Transkripte | Zusatzmaterial ▶ | ▼ Recherche | ▼ Hilfe
AGD Pragmatik IDS
DGD 2.0 - intern > Recherche > Volltext-Suche > Volltext-Suche in Transkripten

Struktur: Korpusauswahl: [alle Korpora (DGD 1.0)] Anfrage:

● Transkripte

- ausgewählt: alle Korpora (DGD 1.0)
- alle Korpora (DGD 1.0)
- DH: Deutsch heute
- DS: Dialogstrukturen
- FR: Grundstrukturen: Freiburger Korpus
- PF: Deutsche Umgangssprachen: Pfeffer-Korpus

Schule

Wildcards: [_] für ein beliebiges Zeichen, [%] für eine Folge beliebiger Zeichen.
Groß-/Kleinschreibung wird beachtet.
Erweiterte Suchoptionen:
Suchbegriffe können mit [AND] bzw. [OR] verknüpft werden:
Italien AND Österreich
oder Suche Mit dem [NEAR]-Operator lässt sich der genaue Wortabstand von Suchbegriffen spezifizieren:
z.B. *...Uraufführung des Freiburg-Films anlässlich der 850-Jahrfeier von Freiburg...*
NEAR((Uraufführung, Jahrfeier), 6, TRUE)
oder *...Lesetext1: Fabel (Nordwind und Sonne), normales Tempo...*
NEAR((Lesetext1, normales Tempo), 4, TRUE)

Trefferliste für den Suchbegriff [Schule] (gefunden in [1551] Transkripten):

1.	FR004.SGM ▶	gingen in die Schmiede sie rissen die Schuljungs aus der **Schule** und tanzten mit ihnen herum also es ist der alte	▶
2.	FR009.SGM ▶	ich möchte am Anfang die Freiherr vom Stein **Schule** auf mehrfache Weise beglückwünschen zu dem was sie heute vorhat	▶
3.	FR018.SGM ▶	es dabei noch einige Anfangsschwierigkeiten aber wir gehen in die **Schule** und es wird eben von Jahr zu Jahr besser und	▶
4.	FR033.SGM ▶	werden wir bis Anfang kommenden Frühjahrs so weit sein die **Schule** in Betrieb zu nehmen und zwar für Nord Baden und	▶
5.	FR047.SGM ▶	ist es dann so sie wollen nach dem Streß der **Schule** ich würde sagen das ist auch ne Art Streß der	▶
6.	FR064.SGM ▶	bei uns in der **Schule** n Lichtbildervortrag	▶
7.	FR067.SGM ▶	Schulleiter weil die dann durch die Kinder die in die **Schule** gehen erfahren können wo noch irgendwie etwas vorhanden ist so	▶
8.	FR071.SGM ▶	als das beste betrachtet hat kann man ruhig aus der **Schule** plaudern so hat sich er entschuppt und ich möchte sagen	▶
9.	FR086.SGM ▶	abhören absehen von vorhin eben gerade und du gehst zur **Schule**	▶
10.	FR103.SGM ▶	sie wenigstens mal den Grund und Boden haben um eine **Schule**	▶
11.	FR113.SGM ▶	Mergelstetten gemacht wir haben s auch in der Friedrich Voith **Schule** schon gemacht des kommt darauf an wo wir halt die	▶

Figure 9: Full-text search in semi-structured transcript data with search results (KWIC-list)

Context sensitive straight to the point queries as in the case of XQuery interfaces are not possible with full-text search against unstructured data. Short query response times and user interfaces that are simple to handle are the main advantages of full-text retrieval strategies.

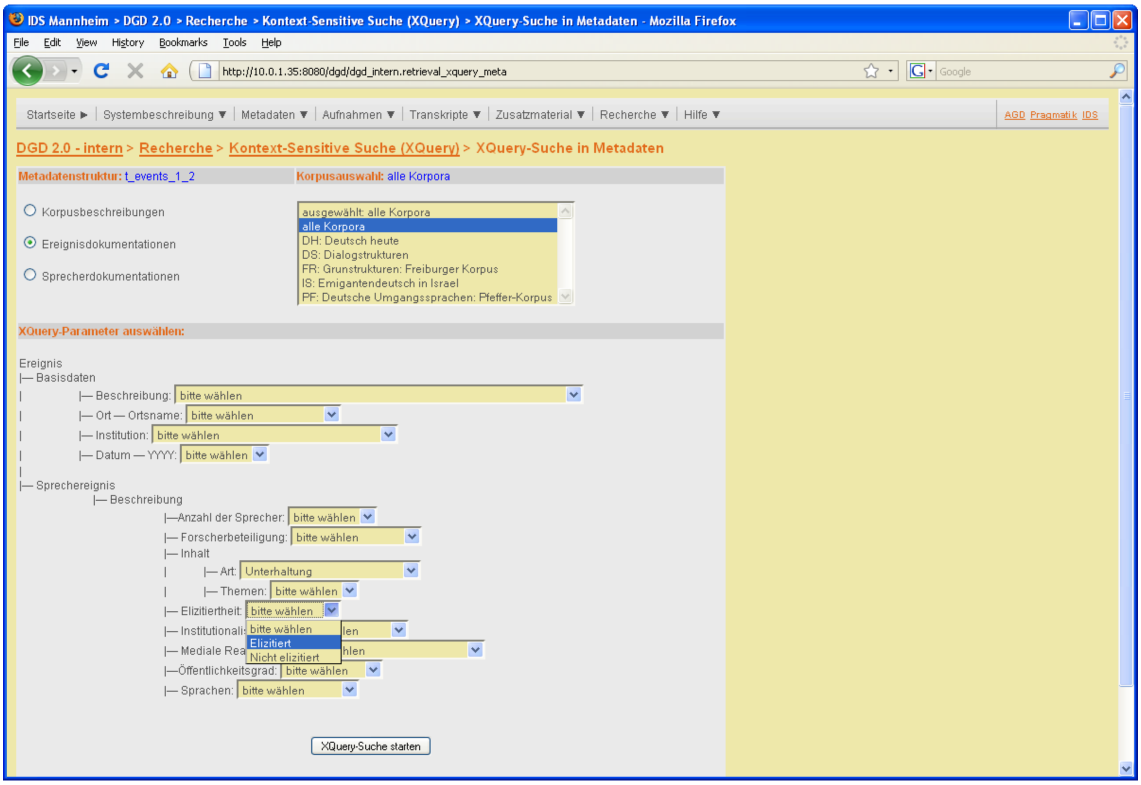
However, an additional full-text option for structured documents (like speech corpus XML meta-documentation) can also be helpful for users that are not familiar with the specific document structure. The disadvantages are obvious: the hits provided by full-text search in structured documents are completely context-free because an exact specification of document sections (like single or complex XML elements) as with the use of XQuery is not possible.

3.2 XQuery Information Retrieval in structured XML Documents

For the information retrieval in fine-grained XML instances like meta data or time aligned multi-dimensional transcripts, the full-text search option without context consideration may not be sufficient. For the hierarchically interdependent informational units of XML structured data a context sensitive retrieval approach is implemented based on the XML query language (XQuery). XQuery allows the definition of complex queries, providing a wide set of XML based retrieval capabilities like for example criteria specific information selection and filtering, joining of data from document selections, sorting, grouping, aggregating, transforming and restructuring of data and arithmetic calculations on numbers and dates[15].

XQuery is a powerful query language operating on XML instances stored in an object-relational XML database, similar as SQL[16] queries are working against relational database tables. In both query cases, a deeper knowledge about the underlying information structure is required for the person who is defining and executing the query. This means for Web-based XQuery retrieval interfaces: two different approaches can be implemented for online XQuery execution.

The first is a HTML form providing a graphical representation of the XML information tree including input fields or listings with possible field contents. After submitting the form, the XQuery syntax code is generated from the HTML form inputs and executed on server side. Figure 10 shows a draft of a graphical XQuery composition interface:


IDS Mannheim > DGD 2.0 > Recherche > Kontext-Sensitive Suche (XQuery) > XQuery-Suche in Metadaten - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://10.0.1.35:8080/dgd/dgd_intern.retrieval_xquery_meta
Google
Startseite ▶ | Systembeschreibung ▼ | Metadaten ▼ | Aufnahmen ▼ | Transkripte ▼ | Zusatzmaterial ▼ | Recherche ▼ | Hilfe ▼
AGD Pragmatik IDS
DGD 2.0 - intern > Recherche > Kontext-Sensitive Suche (XQuery) > XQuery-Suche in Metadaten
XQuery-Suche starten

Metadatenstruktur: t_events_1_2

Korpusauswahl: alle Korpora

○ Korpusbeschreibungen

● Ereignisdokumentationen

○ Sprecherdokumentationen

ausgewählt: alle Korpora
alle Korpora
DH: Deutsch heute
DS: Dialogstrukturen
FR: Grunstrukturen: Freiburger Korpus
IS: Emigantendeutsch in Israel
PF: Deutsche Umgangssprachen: Pfeffer-Korpus

XQuery-Parameter auswählen:

```
Ereignis
|— Basisdaten
|          |— Beschreibung: bitte wählen
|          |— Ort — Ortsname: bitte wählen
|          |— Institution: bitte wählen
|          |— Datum — YYYY: bitte wählen
|
|— Sprechereignis
             |— Beschreibung
                        |—Anzahl der Sprecher: bitte wählen
                        |— Forscherbeteiligung: bitte wählen
                        |— Inhalt
                        |        |— Art: Unterhaltung
                        |        |— Themen: bitte wählen
                        |— Elizitiertheit: bitte wählen
                        |— Institutionali  bitte wählen  len
                        |— Mediale Rea     Elizitiert    hlen
                                           Nicht elizitiert
                        |—Öffentlichkeitsgrad: bitte wählen
                        |— Sprachen: bitte wählen
```

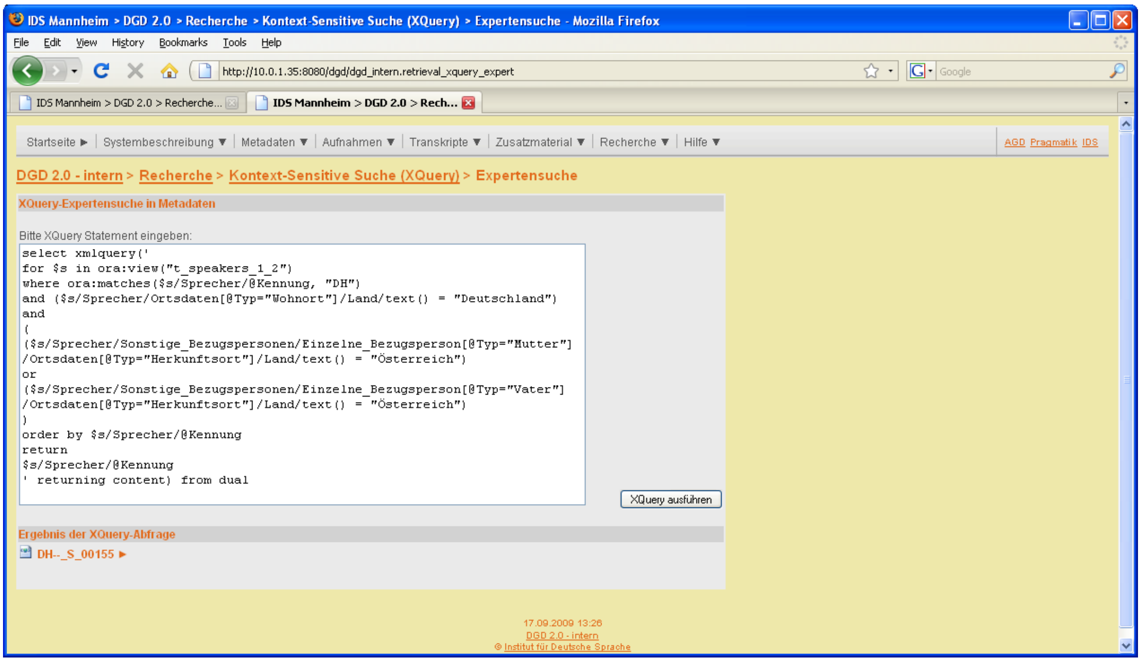
Figure 10: HTML form providing a graphical XQuery composition interface

[15] Cf. Walmsley (2007), p. 1ff.

[16] SQL: the Standard Query Language (relational databases)

XQuery processing via HTML form interfaces may be limited regarding query definition flexibility as it quickly becomes costly in terms of programming effort, especially with queries operating on complex data-centric XML instances or with queries joining distinct XML structures.

A less graphical approach where the XQuery can be edited manually (cf. figure 11) is more suitable to query this type of XML instances. The HTML form provides a text area field where the XQuery syntax can be freely entered as plain text. When the form is submitted, the XQuery input string is directly executed on the database server. This approach is primarily intended for system experts.


IDS Mannheim > DGD 2.0 > Recherche > Kontext-Sensitive Suche (XQuery) > Expertensuche - Mozilla Firefox
File Edit View History Bookmarks Tools Help
http://10.0.1.35:8080/dgd/dgd_intern.retrieval_xquery_expert
Startseite ▶ | Systembeschreibung ▼ | Metadaten ▼ | Aufnahmen ▼ | Transkripte ▼ | Zusatzmaterial ▼ | Recherche ▼ | Hilfe ▼
AGD Pragmatik IDS
DGD 2.0 - intern > Recherche > Kontext-Sensitive Suche (XQuery) > Expertensuche

XQuery-Expertensuche in Metadaten

Bitte XQuery Statement eingeben:

```
select xmlquery('
for $s in ora:view("t_speakers_1_2")
where ora:matches($s/Sprecher/@Kennung, "DH")
and ($s/Sprecher/Ortsdaten[@Typ="Wohnort"]/Land/text() = "Deutschland")
and
(
($s/Sprecher/Sonstige_Bezugspersonen/Einzelne_Bezugsperson[@Typ="Mutter"]
/Ortsdaten[@Typ="Herkunftsort"]/Land/text() = "Österreich")
or
($s/Sprecher/Sonstige_Bezugspersonen/Einzelne_Bezugsperson[@Typ="Vater"]
/Ortsdaten[@Typ="Herkunftsort"]/Land/text() = "Österreich")
)
order by $s/Sprecher/@Kennung
return
$s/Sprecher/@Kennung
' returning content) from dual
```

Ergebnis der XQuery-Abfrage

DH--_S_00155 ▶

Figure 11: HTML form for XQuery plain text submission

In the above expert XQuery example, we look for all speakers of the speech corpus "German Today"[17], currently living in Germany with at least one parent coming from Austria. The query matches one single speaker (id DH--_S_00155) whose father was born in Austria.

4. Summary and Outlook

Media source files become analyzable via their appropriate meta-information. Contemporary speech corpus systems have to close the gap between the processing of binary media data and related meta-information. The need for standardization of speech corpus components is commonly accepted. Nevertheless, the identification of all parameters necessary for a cross-corpus standardization still remains an outstanding goal. Future evolving technologies like the MPEG-7 standard[18] might provide appropriate logic to achieve the standardized integration of the different audiovisual

[17] The speech corpus project "German Today" (corpus id DH) is ongoing. Currently about 600 speaker documentation instances are stored in the XML database.

[18] MPEG-7 ISO/IEC Standard

information types like audio, voice, video, images, graphs and 3D models, potentially involved in media corpora.

References

Gasch, Joachim (2008): XML Schema driven Database Management of Speech Corpus Metadata. In: SDV - Sprache und Datenverarbeitung/ International Journal for Language Data Processing. Vol. 32.1/2008. S. 23-33.

International Organization for Standardization (ISO) (2004): Coding of moving Pictures and Audio, MPEG-7 Overview.
URL (07/2009): http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

Nielsen, Jakob (1994): Usability Engineering, published by Morgan Kaufmann, San Francisco, 1994.

Schneider, Roman (2009): Information Retrieval mit Oracle Text. In: iX - Magazin für professionelle Informationstechnik, Heft 9/2009. S. 144-147.

Trippel, Thorsten (2004): Metadata for Time Aligned Corpora. In: Proceedings of the Workshop "A Registry of Linguistic Data categories within an Integrated Language Resources Repository Area" at the fourth International Conference on Language Resources and Evaluation (LREC), Lisbon, 2004.

U.S. General Services Administration (GSA) (2005): Efficient and Effective Information Retrieval and Sharing (EEIRS) Request For Information (RFI) Response Analysis.
URL (07/2009): http://www.cio.gov/documents/EEIRS_RFI_Response_Analysis.pdf

Walmsley, Priscilla (2007): XQuery, O'Reilly Media Inc., Sebastopol, CA, 2007.

Wichmann, Anne (2008): Speech corpora and spoken corpora. In: Corpus Linguistics, Part 1; An International Handbook; Edited by Lüdeling, Anke; Kytö, Merja; Berlin, New York (Mouton de Gruyter) 2008.

Table of Contents

Table of Figures