memasysco: XML schema based metadata management system for speech corpora

Joachim Gasch, Caren Brinckmann, Sylvia Dickgießer
Institut für Deutsche Sprache
R5 6-13, Mannheim, Germany
E-mail: {gasch, brinckmann, dickgiesser}@ids-mannheim.de

Abstract

The metadata management system for speech corpora "memasysco" has been developed at the Institut für Deutsche Sprache (IDS) and is applied for the first time to document the speech corpus "German Today". memasysco is based on a data model for the documentation of speech corpora and contains two generic XML schemas that drive data capture, XML native database storage, dynamic publishing, and information retrieval. The development of memasysco's information architecture was mainly based on the ISLE MetaData Initiative (IMDI) guidelines for publishing metadata of linguistic resources. However, since we also have to support the corpus management process in research projects at the IDS, we need a finer atomic granularity for some documentation components as well as more restrictive categories to ensure data integrity. The XML metadata of different speech corpus projects are centrally validated and natively stored in an Oracle XML database. The extension of the system to the management of annotations of audio and video signals (e.g. orthographic and phonetic transcriptions) is planned for the near future.

1. Introduction

The IDS (Institut für Deutsche Sprache = Institute for German Language) in Mannheim is the central non-university research institute of German linguistics in Germany. It is one of the major institutions providing German speech and language corpora to linguists. The collection of speech corpora was compiled during the last 50 years and is archived at the AGD (Archiv für Gesprochenes Deutsch = Archive for Spoken German[1]). It contains recordings, transcripts, and supporting documents, mainly from the research areas of dialectology and conversation analysis.

Some of these corpora were compiled within IDS-internal projects but most of them were supplied by external research institutions and archives. Since no universal metadata standard existed at the time of their production, they are documented in a rather heterogeneous fashion. For the corpus-producing projects themselves this is mostly unproblematic, e.g. if a project is mainly interested in syntax, the collection of information about the location or the participants of a speech event might not be very important. However, this practice led to substantial problems regarding the retrieval of language resources and their reusability in different research contexts (cf. Schiel and Draxler, 2004).

Mid-2006 an IDS workgroup was established to find, adapt, or develop a suitable metadata standard for all IDS speech corpora. In this workgroup staff members of the AGD and the speech corpus project "German Today" (Brinckmann et al., in press) cooperate.

Retrofitting existing corpora into a new metadata schema is difficult, especially if some of the needed documentation is not available any more. Nevertheless, the resulting standard will be binding for future IDS speech corpus projects and for the construction of a new comprehensive speech corpus database. The metadata standard and management system must balance project-specific documentations needs with universal applicability. It should allow

- unified validatable documentation in different corpus projects
- project-internal metadata management (in particular validatable data entry and revision ensuring data integrity)
- comprehensive and effective search and retrieval across the metadata of all IDS speech corpora
- publication of corpus metadata.

2. Metadata standards

Developers of metadata standards all face a basic problem: If the proposed standard is too general, it is of little use; if it is too specific, it tends to become burdensome and inflexible. Several standards for the representation of metadata have already been proposed, namely DC (Dublin Core[2]), OLAC (Open Language Archive[3]), TEI (Text Encoding Initiative[4]), MPEG-7[5], and IMDI (ISLE MetaData Initiative[6]). Of these standards, IMDI is most elaborate and specialized for speech corpora intended for linguistic research (see Trippel, 2004), whereas DC and OLAC are very generic.

IMDI provides two separate XML schemas for the representation of metadata: one schema for metadata on the corpus level ("catalogue descriptions") and one on the level of corpus components ("session descriptions"). Metadata about participants (i.e. speakers, researchers, and annotators) are included in the session metadata. Since IMDI standards are mainly intended for corpus publication, this ensures that each session description is independent and can be used on its own. Another aim of IMDI is to provide a metadata standard with only very

[1] URL: http://agd.ids-mannheim.de/

[2] URL: http://dublincore.org/

[3] URL: http://www.language-archives.org/OLAC/metadata.html

[4] URL: http://www.tei-c.org/

[5] URL: http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

[6] URL: http://www.mpi.nl/IMDI/

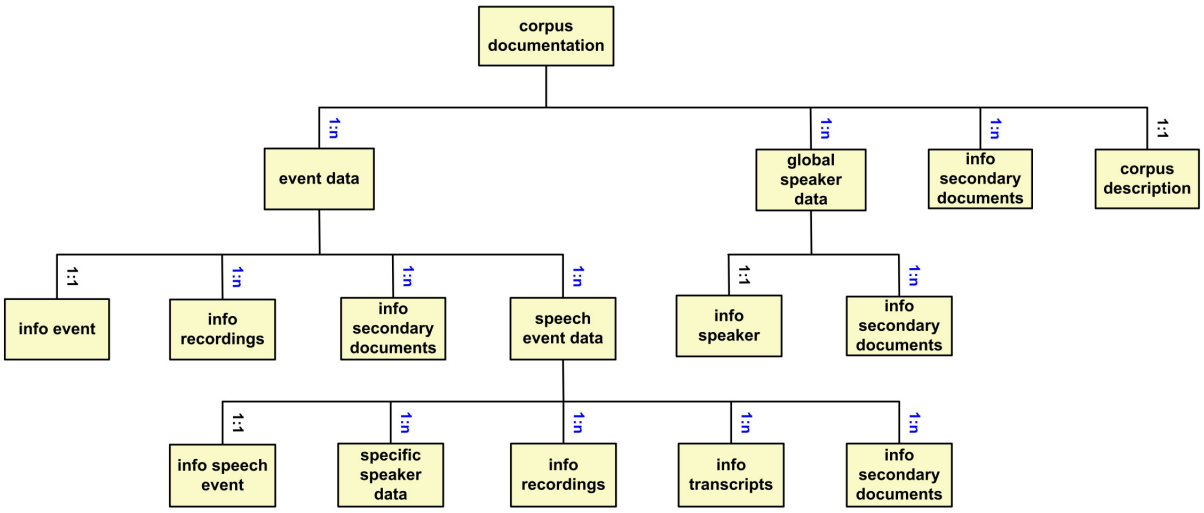

Figure 1: memasysco's data model of corpus documentation

few mandatory elements, thus accommodating to the publication of many different corpora. The session descriptions can be adapted to project-specific needs by adding "keys" in the form of unrestricted name/value pairs.

In a first attempt to provide one metadata standard for all IDS speech corpus projects we examined the IMDI session description schema and detected two problematic features:

1. As the "keys" elements can be filled at will, an adaptation of the schema for project-specific needs by using "keys" would not be restrictive enough to ensure thorough data integrity across the metadata of several different corpus projects.
2. All metadata regarding circumstances, conditions, and participants of one linguistic event are bundled in one schema. This can lead to large redundancies in a database if several linguistic events with identical social contexts have to be documented and if particular individuals take part in several linguistic events.

We therefore decided to develop a model of corpus documentation and two new XML schemas based on existing internal and external metadata structures and recommendations for corpus documentation.

3. memasysco

3.1 Data model and XML schemas

The metadata management system for speech corpora "memasysco" is an instrument for speech corpus documentation at the IDS. It is based on a model of corpus documentation containing four domains that are structured using XML schemas (see Figure 1).

The main differences between this model and the model behind IMDI's session description are the separation of event and speech event data and the fact that event-independent global speaker data are handled by a separate schema. So far, two generic XML schemas were developed and implemented: a schema for the documentation of corpus components regarding events ("event schema") and a schema for global speaker data ("speaker schema"). These two extensive repositories constitute the basis for project-specific schemas. They contain mandatory and optional components, field descriptions, default values, and value restrictions.

For each corpus-building project the general XML schemas are adapted to project-specific needs. The customization is achieved by specifying:

- the number of possible occurrences for each element (e.g. an originally optional element can be made mandatory by setting the minimum number of occurrences to "1")
- facets, i.e. restrictions on XML elements such as enumerations, patterns, and value restrictions
- fixed values, i.e. values that cannot be changed and can therefore be ignored when entering the data
- default values, which are used when a new document instance is generated.

Many of the default values are only used in newly generated document instances to make them instantly valid(atable). They are set to values that are not used in the final documents, e.g. the value "Obligatorisch" (English: "mandatory") for strings. During editing and revision it can easily be checked whether all elements have been filled with appropriate real content before importing an instance into the database.

3.2 Data entry and revision

XML documentation instances are edited using Altova Authentic (Browser Edition)[7], a browser-based schema validating XML editor solution (see Figure 2). The XML editor browser plug-in is driven by a web based client/server application implemented in Oracle PL/SQL.

[7] Altova Authentic XML Authoring, URL: http://www.altova.com/products/authentic/xml_db_form_editor.html

Personalized user access, unique document ID assignment, and document workflow are controlled by this application.

The XML Editor provides an ergonomic support to the user during XML data entry and also guarantees semantic and syntactic data quality due to the customization of the underlying project-specific XML schemas.

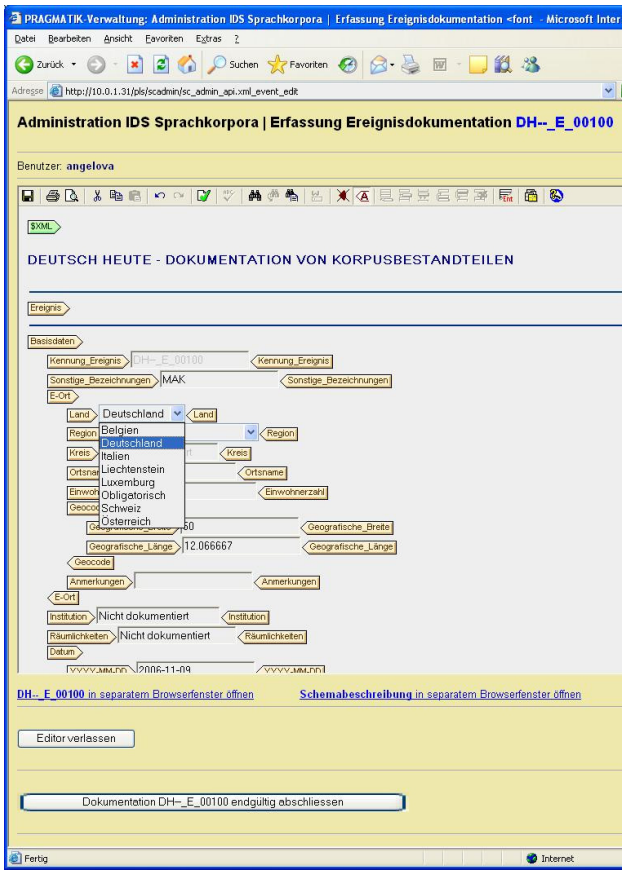

Figure 2: Data entry with the browser-based XML editor

Some examples for XML schema driven content specifications are illustrated in the following paragraphs.

Enumerations

The enumeration content model is used to reach a maximum standardization of field contents. The editor displays the possible element contents as a drop-down list where one unique value has to be selected (see Figure 3).

Regular expressions

Content patterns that are defined for elements in the XML schema are automatically checked in real time while the user is editing the field. The field content is displayed in red if the pattern of the underlying regular expression is violated (see Figure 4).

Mandatory non-empty elements

When a new document instance is generated (using default values), the document is valid. The XML editor permits the user to validate the document at any time against the XML schema to identify possible errors during data entry. E.g. an empty element that is not allowed by the schema would produce an error message when the document is validated.

Multiple elements (also containing child elements)

The XML editor enables the user to duplicate specific complex elements including its substructure tree as defined in the XML schema.

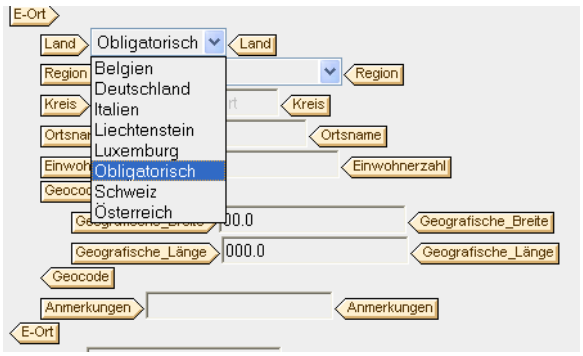

Figure 3: Enumerations are displayed as drop-down lists.

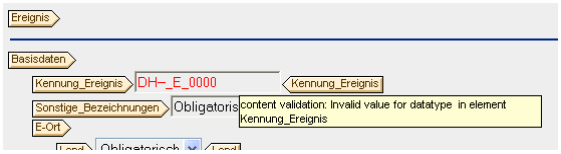

Figure 4: Example of a violated underlying regular expression

3.3 Oracle 11g XML Database

Electronic metadata information of speech corpora can stem from very different sources. Metadata of newly compiled corpora are entered and edited manually with our XML Editor. The complete XML document workflow process including document generation, XML schema based validation, document revision, and XML database storage is centrally managed by the Oracle PL/SQL application framework. The system is logging important work steps while the documents are edited during distinct workflow cycles by different users.

In case of pre-existing electronic metadata we will have to set up an automated bulk data mapping process to validate and import the data into the XML database. For metadata exchange with other institutions an automatic mapping to other standards such as IMDI is also possible.

3.3.1 XML Data Storage

Once the XML instances have been finally revised, the documents are ready for import into the XML database. The Oracle object relational XML database[8] basically distinguishes two different types of internal processing and storage of XML instances: unstructured vs. structured database storage.

[8] Oracle XML DB, URL: http://www.oracle.com/technology/tech/xml/xmldb/index.html

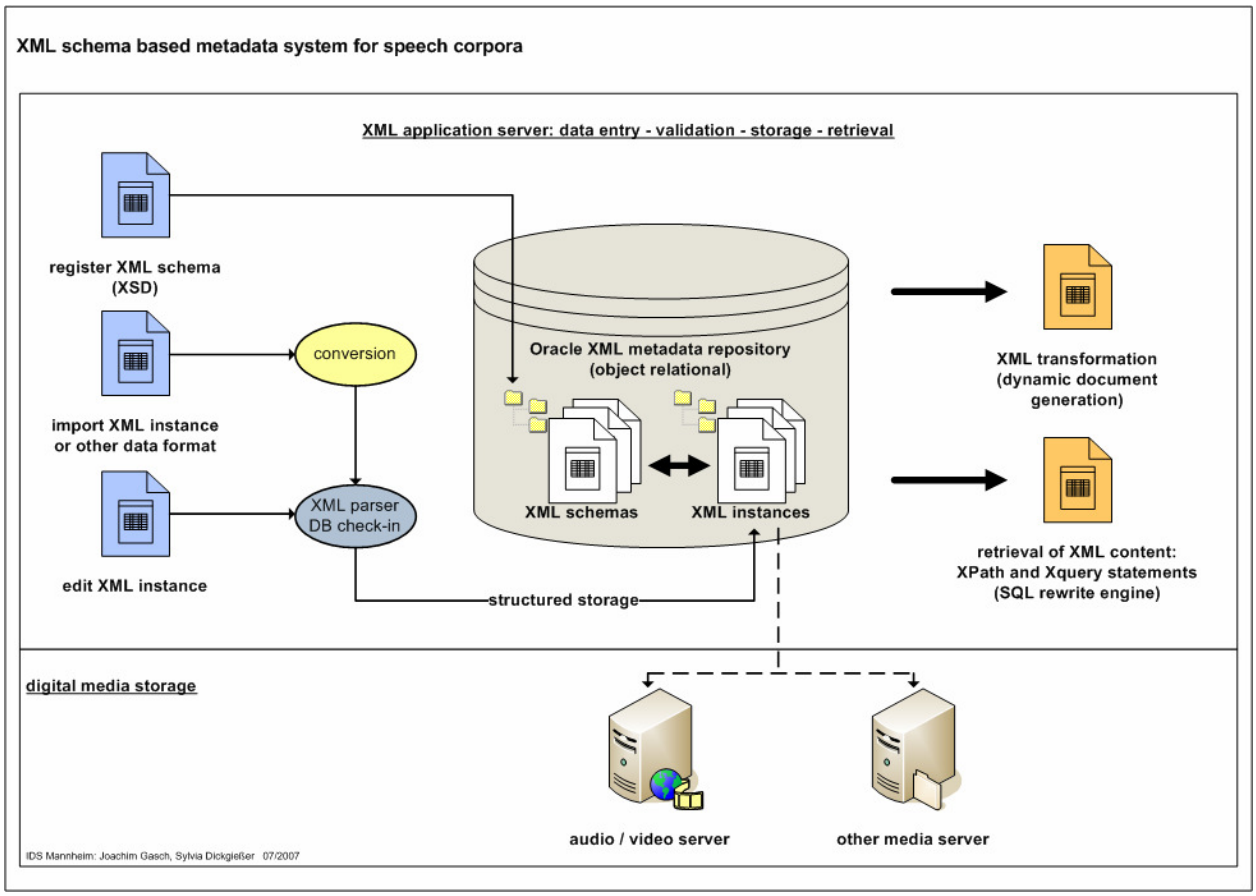

Figure 5: XML application server: data entry - validation - storage - retrieval

In the unstructured case, the XML database only checks if a document is syntactically well formed without validating it against an XML schema and writes the document instance into an XMLType CLOB field. When retrieving document content, the Document Object Model (DOM) has to be built first for each XML instance. This makes this alternative very time consuming for large amounts of data and complex XML structures.

We decided to use the structured storage approach (schema aware processing of the XML data) to store the documentation instances of our project (see Figure 5). With this approach, first the XML schema is registered in the database and an XMLType table is generated according to the XML schema. Then an internal object relational database structure is created modeling the DOM of the XML instances as described in the XML schema. The XML instances are semantically validated against their XML schema during the import process (XML parser). Using the object relational approach, content retrieval is much faster than with unstructured storage because the DOMs do not have to be built at runtime.

3.3.2 Information Retrieval

The XML schema aware processing of XML instances enables the implementation of context sensitive query interfaces to retrieve informational units including their semantic contexts. In the relational database world, the Standard Query Language (SQL) is used to retrieve information from relational table structures. Within XML technology, XQuery[9] (as an extension of XPath) is the W3C recommendation providing the correspondent navigation and retrieval functionalities to extract information from complex XML structures.

Figure 6 shows the result screen for a complex XQuery example: we select the speech event IDs, the town names and the geocodes for all speech events that

- have been recorded in Germany
- in the federal states of “Baden-Württemberg” or “Sachsen” and
- belong to the speech corpus “DH” (“Deutsch heute” = “German Today”).

3.3.3 Online Publishing

The bundled XML database storage of all documentation components of speech corpora on corpus, event, and global speaker data level allows the publication of personalized views of browsable, well structured corpus information. Such fine grained views may vary depending on individual user needs. They are dynamically generated via XSL transformations operating on the underlying XML data collections.

Figure 7 shows an example of a customized view of the event documentation. The result window shows all subsequent documentation information for the complex element node “Quellaufnahme” (English “original re-

[9] W3C, XQuery 1.0: An XML Query Language, URL: http://www.w3.org/TR/xquery/

cording"). The media resource (wav file) is dynamically linked with the documentation view during the XSL transformation process.

4. Outlook

For the near future we plan to extend the system to include the management of XML schema based annotations of audio and video signals (e.g. orthographic and phonetic transcriptions). The integration of meta information and signal annotations into the XML database storage model will allow us to bundle all corpus related information for the retrieval processes and to perform complex queries against the complete XML data collections.

5. Acknowledgements

We thank Ralf Knöbl and Bistra Angelova of the corpus project "German Today" for their helpful feedback during the ongoing development of memasysco. We also wish to thank three anonymous reviewers for their valuable comments on an earlier draft of this paper.

6. References

Brinckmann, C., Kleiner, S., Knöbl, R., Berend, N. (in press). German Today: an areally extensive corpus of spoken Standard German. In *Proceedings of the sixth international conference on Language Resources and Evaluation, LREC 2008*, Marrakech, Morocco.

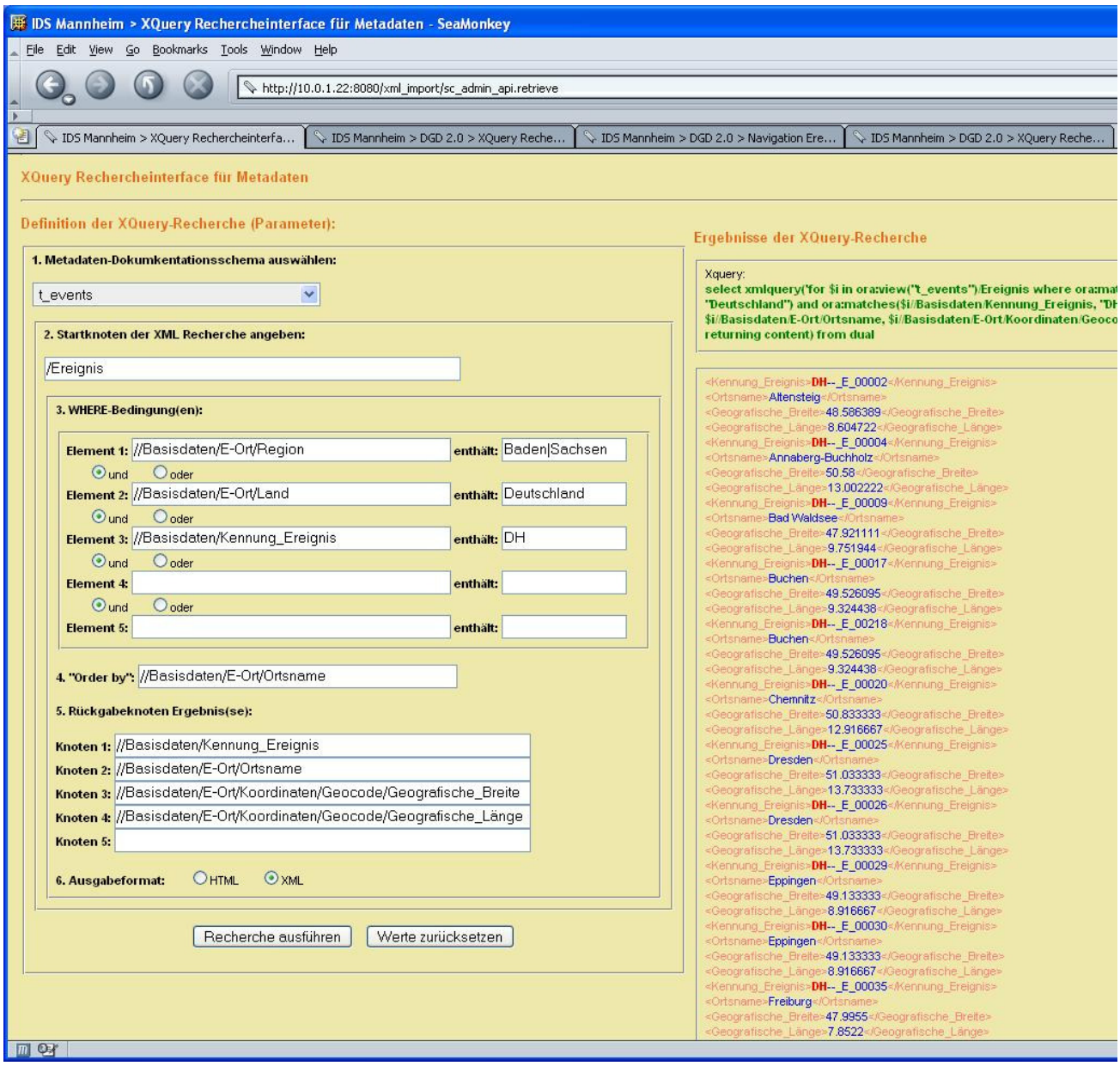

Figure 6: XQuery retrieval interface for event documentation

Trippel, T. (2004). Metadata for time aligned corpora. In *Proceedings of the workshop: A registry of linguistic data categories within an integrated language repository area, LREC 2004*, Lisbon, Portugal. URL: http://coral.lili.uni-bielefeld.de/~ttrippel/papers/multimodal_metadata_rev2.0.pdf (accessed March 28, 2008)

Schiel, F. and Draxler, C. (2004): *The production of speech corpora. Version 2.5*. URL: http://www.bas.uni-muenchen.de/Forschung/BITS/TP1/Cookbook/ (accessed March 28, 2008)

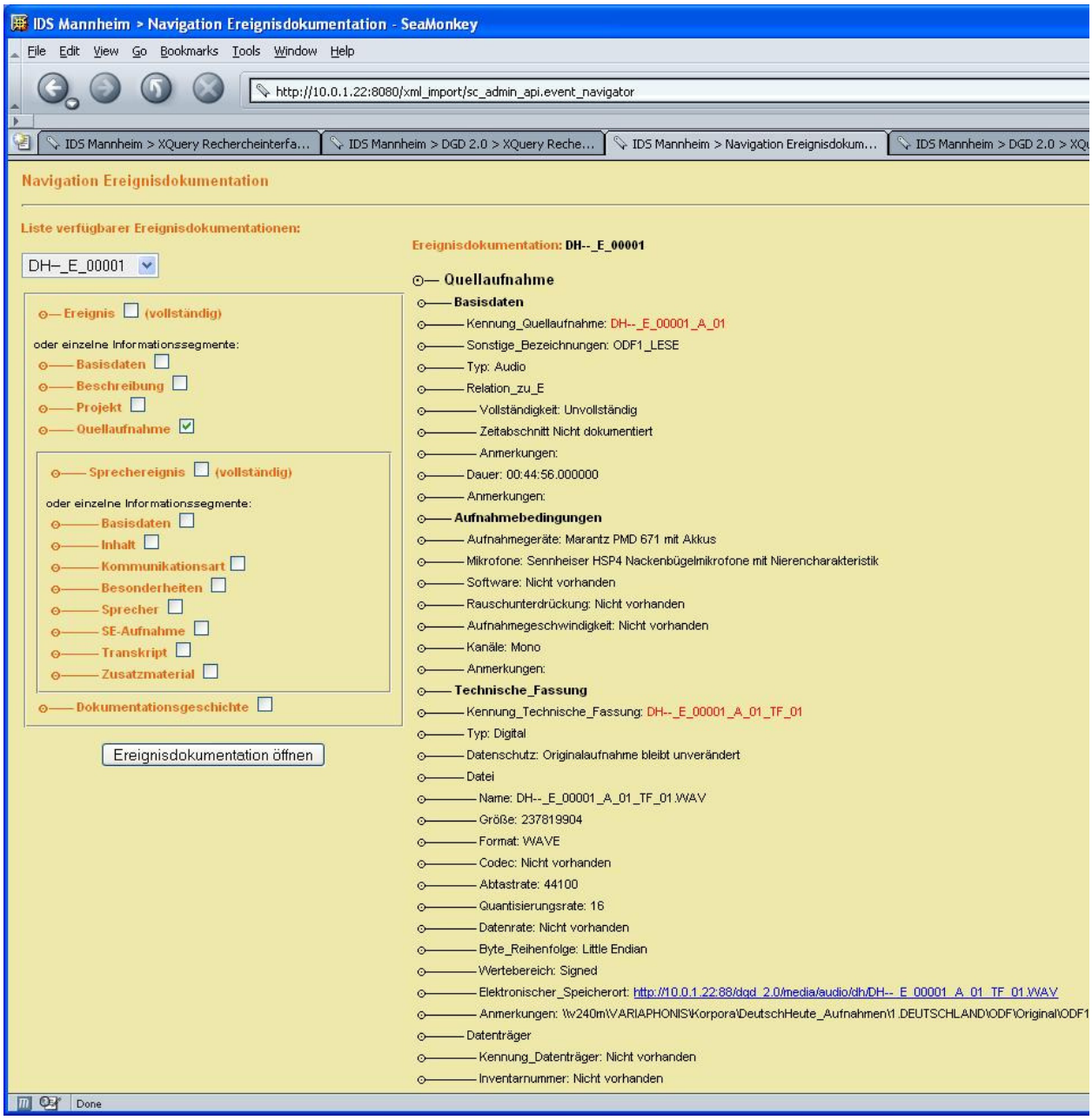

Figure 7: Customizable navigation of event documentation