Minimally Redundant Metadata Schemas for Speech Corpora

Caren Brinckmann,[1] Sylvia Dickgießer
and Joachim Gasch

Abstract

We present an XML-based metadata standard for the documentation of speech and multimedia corpora that was developed at the Institute for German Language (IDS) in Mannheim, Germany. The IDS is one of the major institutions providing German speech and language corpora to researchers. These corpora stem from many different sources and were previously documented in a rather heterogeneous fashion using a variety of data models and formats.

In order to unify the documentation for existing and future corpora, the IDS-internal Archive for Spoken German collaborated with several projects and developed a set of standardised XML metadata schemas. These XML schemas build on existing internal and external documentation schemas (such as IMDI) and take into account the workflow of speech corpus production. In order to minimise redundancy, separate schemas were designed for projects, speakers, recording sessions, and entire corpora.

The resulting schemas are tested in ongoing speech and multi-media projects at the IDS and are regularly revised. They are accompanied by element definitions, guidelines, and examples. In addition, a mapping to IMDI will be provided.

[1] *e-mail*: brinckmann@ids-mannheim.de