Sociolinguistics metadata tags

Metatdata Tags Frequently used by Sociolinguists

Ideally, every audio or video file generated as part of a research project should include data about the data ("metadata"). Metadata is used to facilitate the understanding, use and management of data. The metadata required for this will vary with the type of data and context of use. In sociolinguistics, there is a common core of sociodemographic information that is of interest for addressing particular types of research questions related to language change and variation. For this reason, it seems reasonable to publish a common set of data that researchers may choose to associate with recordings generated in field research. Supplying information in datafields not of immediate interest to a research study is worth the effort because it carries a minimal time burden, but greatly increases the usability of data for other researchers and types of study.

In its simplest form, metadata may be stored in a text file (.txt or .rtf) that sits in the directory or folder with the sound or video file. For metadata attached to Web pages, the standard encoding scheme is HTML (HyperText Markup Language). RDF (Resource Description Framework) supports multiple metadata schemes.

Below is a list of metadata tags used at the University of Washington, Sociolinguistics Laboratory. The recording metadata categories below are also the basis for a digital CD archive that drives a CD carousel. Fields in the INDEX of the CD Archive are searchable, and CDs in the catalogue may be located and automatically ejected using search criteria.

AUDIO RECORDING INFORMATION

Name of audiorecording file(s)
    transcription or annotation filename(s)
    supplemental filename(s)
Date recording made
Location of recording
Format of sound or video file (.wav, .aiff)
Sampling Rate
Microphone model
Recording device

Project name
Project website (url)
Name of data administrator or investigator
Data administrator contact information
IRB approval number

Publications associated with the project (appropriate for bibliographic citation)

Register (or genre)
Type of recorded data (unscripted conversation, dyadic or small group interview, individual interview, reading passage, wordlist, minimal pair list, words in isolation, self-commutation test, map task, attitude or subjective reaction test)
Names of elicitation instruments used to elicit data (with filename, as appropriate)
Version history available for datafiles?

PHONETIC ANALYSIS TAGS

Analysis Window length
Analysis poles
Analysis scripts used
Normalization method

SPEAKER-LEVEL TAGS

Speakers on recording (copy below as needed or insert table)
For each speaker, include:

- Name
- Sex
- Age
- Age cohort
- Known speech impediments or disorders
- Ethnicity
- Socioeconomic class
- Highest educational level attained
- Occupation
- Place of birth
- Residence history (places lived for more than 6 months)
- Regionality
- Social Network information available (yes/no)? If yes, name of datafile:
- Neighborhood
- Language background (all language varieties [dialect region/language name] spoken)
- Bi/Multilingual (yes/no)
- Languages spoken natively
- Languages of high fluency
- Languages of low fluency
- Writing system used or preferred by speaker
- Level of literacy

GROUP-LEVEL TAGS

Language
Modality (if signed language)
Dialect
Task (wordlist, reading passage, casual conversation, etc.)

Bi/Multilingual or dialectal community (yes/no)

TOKEN-LEVEL TAGS

Vowel (IPA category)
Word
Preceding phone
Following phone
Place
Manner
Voicing
Phonation type
Normalized (y/n)
Stress (primary/secondary/unstressed)
Tone level