6. Data Management Plan

The programs resulting from this project will be registered as open-source projects at sourceforge.net, as well as the LREC language resources repository. The textual data will be stored in an extended version of the TalkBank format[1] that will also be compatible with the ANNIS3 framework. This will allow us to include the data in TalkBank, ANNIS, and other open repositories such as the Virtual Linguistic Observatory (VLO) or the Linguistic Data Consortium (LDC).

Consistency of data format will be achieved by use of the TalkBank *Chatter* program that checks XML data for validity against the schema and outputs text formats in the required display formats. To further guarantee validity, it creates a round trip in which text in the display format is reprocessed into XML and compared against the original to make sure that nothing has been altered during formatting and that all codes are accurately stored in the XML archive. TalkBank also periodically runs custom Unix-based tools for automatic checking of the overall database for annotation linkage, integrity of directory structure, proper harvesting of metadata, mirroring of data to other websites, and availability of resources. All methods are part of the process of obtaining the Data Seal of Approval for TalkBank described at http://talkbank.org/share/preservation.html and http://talkbank.org/share/workflow.html.

[1] http://talkbank.org/software/talkbank.xsd