Implementation of the Data Seal of Approval

The Data Seal of Approval board hereby confirms that the Trusted Digital repository TalkBank complies with the guidelines version 2014-2017 set by the Data Seal of Approval Board.
The afore-mentioned repository has therefore acquired the Data Seal of Approval of 2013 on April 8, 2014.

The Trusted Digital repository is allowed to place an image of the Data Seal of Approval logo corresponding to the guidelines version date on their website. This image must link to this file which is hosted on the Data Seal of Approval website.

Yours sincerely,

The Data Seal of Approval Board

Assessment Information

Guidelines Version:	2014-2017 | July 19, 2013
Guidelines Information Booklet:	DSA-booklet_2014-2017.pdf
All Guidelines Documentation:	Documentation
Repository:	TalkBank
Seal Acquiry Date:	Apr. 08, 2014
For the latest version of the awarded DSA for this repository please visit our website:	http://assessment.datasealofapproval.org/seals/
Previously Acquired Seals:	None
This repository is owned by:	**Carnegie Mellon University** 254M Baker Hall CMU - Psychology 5000 Forbes Avenue 15213 Pittsburgh PA USA T +1 412 268-3793 E macw@cmu.edu W http://talkbank.org/

Assessment

0. Repository Context

Applicant Entry

Self-assessment statement:

A General Description of TalkBank

TalkBank is an archive of transcripts of spoken language interactions, many of which are linked to either audio or video. The major designated communities involved include child language researchers, aphasiologists, linguists, conversation analysts, and second language acquisition researchers. Long-term data preservation is provided by Carnegie Mellon University and CLARIN (www.clarin.eu). Several of the CLARIN centers have received the Data Seal of Approval and TalkBank data is currently mirrored by the CLARIN Center at the MPI in Nijmegen that has the Data Seal of Approval. The only outsourcing we do is for data mirroring to guarantee preservation. This project has been funded continuously by the National Institutes of Health since 1984 and has also received support from the National Science Foundation and the MacArthur Foundation. A search of scholar.google.com shows that there are now 4350 published articles based on use of the TalkBank databases. Current NIH support involves three major ongoing five-year grants for child language, aphasia, and phonology. The central website is http://talkbank.org. Within the overall TalkBank corpus, there are several subcorpora, the largest and oldest of which is CHILDES (Child Language Data Exchange System) located at http://childes.talkbank.org .

In the responses to the Guidelines, “we” refers to the programming and data analysis staff employed by the TalkBank Project at Carnegie Mellon. The term “producers” refers to the scholars who contribute data. The term “users” refers to the scholars who use the data. All URLs were visited on Tuesday January 7th, 2014.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

1. The data producer deposits the data in a data repository with sufficient information for others to assess the quality of the data, and compliance with disciplinary and ethical norms.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

Data producers are the scholars who have collected the spoken interactions and produced the transcripts and media that are then included in TalkBank. We support the data producers and guarantee data quality through these methods:

1. We teach best practices through online tutorials available at http:://childes.talkbank.org/tutorial.zip that explain how to use the CLAN program for transcription and analysis.

2. We conduct seminars and tutorials at international meetings of the relevant professional societies.

3. The CLAN program is openly downloadable for free at http://childes.talkbank.org/clan/

4. The CLAN manual is freely downloadable from http://childes.talkbank.org/manuals/CLAN.pdf.

5. The data producer verifies the correctness of the transcription process using the CHECK command inside CLAN and the XML checker available from http://talkbank.org/software/chatter.html. After contribution, we

6. Methods for data submission are found at http://talkbank.org/share/contrib.html

7. Standards for corpus documentation are presented in chapter 4.5 of the CHAT manual at http://childes.talkbank.org/manuals/CHAT.pdf

8. Data producers/contributors provide data release forms as found at http://talkbank.org/share/irb/release.pdf

9. Adherence to ethical norms is treated through the IRB (Instiutional Review Board) process summarized at http://talkbank.org/share/irb/

10. Metadata regarding each file is constructed in accord with the CMDI standard at http://www.clarin.eu/content/component-metadata and is included in each archive.

11. We track citations of corpora and the database through yearly reviews using scholar.google.com, as well as letters from contributors.

Corpora that meet all of these standards are judged to be valuable and are included in TalkBank.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

2. The data producer provides the data in formats recommended by the data repository.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

1. The repository has created a manual for the single required format. The format is called CHAT and the manual is available at http://childes.talkbank.org/manuals/CHAT.pdf. There are also published translations into Japanese, Italian, Portuguese, Chinese, and Spanish, as well as several introductions and tutorials.

2. Quality control is achieved by running the CHECK command inside the CLAN program and then the XML checker available at http://talkbank.org/software/chatter.html .

3. The tools used to guarantee correct use of the CHAT standard are the CHECK command and the XML checker described in (2) above.

4. We do not accept any data that are not in CHAT.

5. We do not require detailed statements about data formats, because all data must be in CHAT format.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

3. The data producer provides the data together with the metadata requested by the data repository.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

1. Using the documentation provided by producers, we create metadata files for each resource. For the purposes of harvesting by OLAC (Online Language Archiving Community at http://www.language-archives.org, we produce a single metadata file for each corpus that is included in the relevant .zip file that can be downloaded. For harvesting in the IMDI/CMDI framework at http://www.clarin.eu/content/component-metadata, we use a program built into CLAN to automatically generate metadata records for each transcripts and media file. These can be seen at http://talkbank.org/data-imdi/talkbank/ and http://childes.talkbank.org/data-imdi/childes/

2. We do not require data producers to generate these OLAC and IMDI/CMDI metadata files. We do this using the data they provide.

3. We enforce a quality check as we create these files.

4. Our metadata formats are in compliance with the two major standards for linguistic metadata documentation, e.g. OLAC and CMDI. Both include Dublin Core as subsets.

5. The primary use of metadata is for resource discovery through OLAC and IMDI. Secondary analysis depends on use of the CLAN programs themselves.

6. It is possible that data producers will have failed to collect or transcribe some data that will turn out in the future to be important. However, because we have the raw media for most of our new corpora, transcriptions can be refined later on. None of these issues should lead to problems in terms of long-term preservation.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

4. The data repository has an explicit mission in the area of digital archiving and promulgates it.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Applicant Entry

Statement of Compliance:

3. In progress: We are in the implementation phase.

Self-assessment statement:

1. TalkBank is a government-funded project at Carnegie Mellon University. The mission of TalkBank is to provide a preservable archive of publicly shared data on spoken language. To support the goal of preservation of government-funded data, including TalkBank, the University, acting primarily through Carnegie Mellon Libraries, is developing a policy for LTDP of resources created at the University, including TalkBank. The initial shape of this policy was approved at Faculty Senate on January 22, 2014 and then approved by Subra Suresh, the new President of the University, who was formerly Director of the National Science Foundation. Details regarding this policy will be finalized by April, 2014.

2. In addition to preservation at Carnegie Mellon, all TalkBank materials are included in the Nijmegen Max-Planck archive at mpi.nl and the CLARIN archives. The plans for CLARIN long term preservation are described at http://clarin.eu/contnent/mission. Brian MacWhinney, the Director of TalkBank, is the Chairman of the CLARIN Scientific Advisory Board. CLARIN is currently adding TalkBank as one if its “A-Level” data repository centers. In that process, CLARIN has requested that TalkBank attain the Data Seal of Approval. When the current director, Brian MacWhinney, retires in 2022, the current Director of the AphasiaBank Project, Davida Fromm, will assume the role of Director.

3. The Mission Statement in regards to LTDP is further implemented by inclusion of all TalkBank data and media inside LDC (The Linguistic Data Consortium, ldc.org), CLARIN through the Nijmegen MPI (http://corpus1.mpi.nl/ds/imdi_browser/ -- see "mirrored corpora"), and the mirror site in Antwerp http://www.clips.ua.ac.be/childes/.

4. TalkBank depends on a deep level of commitment from its component research communities. For child language, aphasia, bilingualism, and CA (Conversation Analysis), this involves maintenance of mailing lists, help centers, presentations at conferences, publications of results in special issues, and summer workshops.

5. We do not outsource our basic functions. However, we mirror data through the CLARIN Max Planck Institute Center in Nijmegen that has the DSA (see list of DSA Seals at assessment.datasealofapproval.org).

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

5. The data repository uses due diligence to ensure compliance with legal regulations and contracts including, when applicable, regulations governing the protection of human subjects.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

1. The repository is supported by Carnegie Mellon University, which is the relevant legal entity in contractual matters.

2. We use a standard data contribution form given at http://talkbank.org/share/irb/release.pdf

3. Data consumers are asked to follow our usage guidelines as stated at http://talkbank.org/share/

4. Our conditions and terms of used are given at http://talkbank.org/share/

5. If conditions are not met, we make cases of non-compliance known to the research community. In the 28 years of functioning of TalkBank and CHILDES, there has never been a case of non-compliance.

6. We insure compliance with national and international laws through the IRB (Institutional Review Board) procedure at Carnegie Mellon University. Copyright is based on a Creative Commons License declared at the bottom of the homepage.

7. Data with disclosure risk are password protected. All data in AphasiaBank are password protected. About 3% of the data in other areas are in this category. This is explained in detail at

http://talkbank.org/share/irb/options.html

8. Data with disclosure risk are password protected.

9. Data with levels of disclosure risk beyond that of password protection are archived but not distributed.

10. Files are anonymized through replacement of lastnames with the word LastName and replacement of addresses with the word Address.

Issues relating to disclosure risk are discussed in detail between the Director and the Data Producer

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

6. The data repository applies documented processes and procedures for managing data storage.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Applicant Entry

Statement of Compliance:

3. In progress: We are in the implementation phase.

Self-assessment statement:

1. TalkBank has a preservation policy based on mirroring in other data sites and longterm preservation by the University. This policy is at http://talkbank.org/share/preservation.html.

2. The data is backed up through github.com and a series of three complete image backups on 3TB thunderbolt disks. Image backups, using ChronSync, are updated weekly using rotation.

3. Data recovery is from the image backups.

4. Risk management is based on trying to minimize the possibility of either complete data loss through disk failure or hacking or partial data loss through system error. The former is addressed through keeping multiple image copies and the latter through running of ChronoSync comparison between image copies and the current archive.

5. Consistency across archival copies is achieved through use of ChronoSync.

6. One image copy is kept offsite, one in another University building, and one in another part of Baker Hall. All are under lock and key.

7. Because the storage media are hard drives, deterioration means disk failure. If one drive fails, we can restore the data from one of the three remaining complete copies. Because these drives are only running during the

copying process, they never fail. The chances of all four failing at once are extremely low. The major possible danger would be catastrophe to the entire city of Pittsburgh. In that case, copies would still be preserved in Nijmegen and throughout European CLARIN centers. Of course, there could also be global catastrophes.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

7. The data repository has a plan for long-term preservation of its digital assets.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

Applicant Entry

Statement of Compliance:

3. In progress: We are in the implementation phase.

Self-assessment statement:

1. Our basic file format relies on text-only Unicode files. We expect only minor changes in this format over time. More importantly, the CHAT coding system continually undergoes changes. To guarantee preservation of the data on this level, we use the Chatter program to make sure that the XML version of the CHAT files can be roundtripped from CHAT to XML and back without changes. Obsolescence of media files is a more difficult problem. For audio, we maintain both MP3 and WAV formats, in hope that the latter could be converted without loss to any new popular formats. For video, we have stored raw video for some corpora, but for others we only have resources to store compressed versions. For those we focus on making sure that everything is in .H264 format.

2. The transcript files will be usable in their current format as long as computers can read text files and Unicode. We have developed programs that convert when necessary to six other current file formats, but we rely on CHAT format as the current standard in the field.

3. These principles are posted at http://talkbank.org/share/preservation.html

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

8. Archiving takes place according to explicit work flows across the data life cycle.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

1. The workflow for the inclusion of data in TalkBank (http://talkbank.org/share/workflow.html) has the following steps: contributors read the guidelines, they send mail to macw@cmu.edu, we reply, data is transferred, we check the data using CHECK and Chatter, we create metadata files, we add documentation to the database documentation files, we create streaming media, we commit all files to github, and we announce the availability of the new corpus on googlegroups mailing lists.

2. We change archival data for two purposes. The first is to update the syntax of coding symbols. This does not involve data loss. The second is to normalize spellings for morphosyntactic analysis. In this case, we maintain the original form alongside the normalized form in the text, using a special code.

3. Our three programmers all have M.S. degrees in Computer Science. Our transcribers are selected for ability to accurately transcribed speech.

4. All data types go through the same workflow.

5. All data are included that conform to the CHAT transcription standard.

6. We never receive data that do not conform to the mission, because the work of putting data into CHAT format already ensures that researchers want to have their relevant data included in the database.

7. Privacy of subjects is guarding through anonymization.

8. Data producers are asked to check over the final versions of their files to make sure they conform to their expectations.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

We would recommend the preparation of procedural documentation (ideally shared online) that reflects these procedures

9. The data repository assumes responsibility from the data producers for access and availability of the digital objects.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

1. We obtain contribution forms from data producers, as given at http://talkbank.org/share/irb/release.pdf

2. We enforce licenses through these contribution forms.

3. Our crisis management plan focuses on possible data loss, as described for Guideline 6.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

10. The data repository enables the users to discover and use the data and refer to them in a persistent way.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

1. Our data are provided in the format required by our communities. In addition, some researchers wish to study transcripts in ELAN format, as described at http://tla.mpi.nl/tools/tla-tools/elan/ , and we can convert automatically from CHAT to ELAN using CLAN.

2. The repository can be fully searched using the commands built into the TalkBank Browser window, as well as through the WebData command inside the CLAN program. OAI harvesting is done through OLAC and IMDI, as described above.

3. We generate PIDs through the CLARIN CMDI system, as described at http://hdl.handle.net/1839/00-DOCS.CLARIN.EU-77

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

11. The data repository ensures the integrity of the digital objects and the metadata.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

1. Files are downloaded from the database in .zip format. Therefore, data loss through transfer will be revealed when the user unzips the file.

2. Integrity of the data is monitored through daily running of XML roundtrip validation, use of ChronoSync for file comparison on image copies, and placement of an @End code at the end of each transcript file.

3. Once data are in the database, we avoid creating multiple versions of data files. However, when changes are made, older versions are stored in github.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

12. The data repository ensures the authenticity of the digital objects and the metadata.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

1. When there are any changes in data format, we advise data producers through an email to our members GoogleGroups list.

2. Files are grouped into corpora that all maintain the same provenance and no additional data is inserted. After data have been included in the database, we maintain an audit trail of changes in terms of github versions.

3. We do not maintain links to other datasets. We maintain links to our own metadata files, as described earlier.

4. When making changes to files, we compare the original with the revised version using DIFF.

5. We maintain complete contact information and personal contacts with all depositors.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

13. The technical infrastructure explicitly supports the tasks and functions described in internationally accepted archival standards like OAIS.

Minimum Required Statement of Compliance:

3. In progress: We are in the implementation phase.

Applicant Entry

Statement of Compliance:

3. In progress: We are in the implementation phase.

Self-assessment statement:

1. The repository is structured in accord with OAIS standards.

2. OAIS standards are implemented in terms of our policies for long term preservation, data migration, metadata creation, storage practices, documentation, legal responsibilities, and data access.

3. Our longterm plan for infrastructure development is to continually expand the scope of the archive and the analysis programs to deal with all aspects of human language. We rely on organization of the relevant research community in each case to make a case for federal funding for each of these separate initiatives.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

We recommend the provision of supporting documentation for this item before the next DSA submission

14. The data consumer complies with access regulations set by the data repository.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

1. We do not require End User licenses.

2. We have no special requirements.

3. We do not need contracts for data access.

4. We use a Creative Commons license CC BY-NC-SA 3.0., as stated at the bottom of our homepage.

5. As stated under Guideline 5, if conditions are not met, we make cases of non-compliance known to the research community. In the 28 years of functioning of TalkBank and CHILDES, there has never been a case of non-compliance.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

15. The data consumer conforms to and agrees with any codes of conduct that are generally accepted in the relevant sector for the exchange and proper use of knowledge and information.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

1. We have a stated codes of conduct policy at http://talkbank.org/share/. Please refer specifically to points 1, 3, and 6.

2. We must deal with Human Subjects. We have received IRB (Institutional Review Board) clearance from Carnegie Mellon and this has been approved by NIH (National Institutes of Health). Contributors also receive IRB review at their institutions.

3. Users agree to the terms at http://talkbank.org/share/.

4. Institutional bodies are not involved in agreeing to the terms of use.

5. As stated under Guideline 5, if conditions are not met, we make cases of non-compliance known to the research community. In the 28 years of functioning of TalkBank and CHILDES, there has never been a case of non-compliance.

6. We provide guidance in the responsible use of all data, as described at http://talkbank.org/share/ particularly points 1, 3, and 6.

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments:

16. The data consumer respects the applicable licences of the data repository regarding the use of the data.

Minimum Required Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Applicant Entry

Statement of Compliance:

4. Implemented: This guideline has been fully implemented for the needs of our repository.

Self-assessment statement:

We rely on Creative Commons license CC BY-NC-SA 3.0 as noted on our homepages. To control password access to data in the AphasiaBank segment of TalkBank with a possible disclosure risk, we rely on three security measures:

1. Password access is only given to fulltime faculty or clinicians with SLP (Speech and Language Pathology) certification from ASHA (the American Speech and Hearing Association). Students can only access data under faculty supervision.

2. Faculty must apply for membership in AphasiaBank and state their intended use of the data.

3. Members agree to the Ground Rules given at http://talkbank.org/share/

Reviewer Entry

Accept or send back to applicant for modification:

Accept

Comments: