TalkBank

Notes Before Completing the Application

We have read and understood the notes concerning our application submission.

True

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background & General Guidance

Glossary of Terms

BACKGROUND INFORMATION

Context

R0. Please provide context for your repository.

Repository Type. Select all relevant types from:

Domain or subject-based repository, Research project repository

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of Repository

TalkBank is an archive of transcripts of spoken language interactions, many of which are linked to either audio or video. Long-term data preservation is provided by Carnegie Mellon University and CLARIN (http://www.clarin.eu) .
CMU-TalkBank is a B-Centre member of the European CLARIN federation; it is the only member of CLARIN outside of Europe, as the map (https://talkbank.org/styles/map.png) shows. TalkBank data is mirrored by the NL-CLARIN Center at the MPI in Nijmegen that has the Core Trust Seal. The 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 http://scholar.google.com shows that there are 8450 published articles based on use of the TalkBank databases. TalkBank websites have received over 8 million hits. Current NIH support involves four major ongoing five-year grants for child language (https://childes.talkbank.org),
aphasia (https://aphasia.talkbank.org), fluency (https://fluency.talkbank.org), and phonology (https://phonbank.talkbank.org). The central website is http://talkbank.org. Within the overall TalkBank corpus, there are several subcorpora, the largest and oldest of which is CHILDES (https://childes.talkbank.org, Child Language Data Exchange System).
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.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Brief Description of the Repository’s Designated Community.

The repository’s designated community: TalkBank provides resources for all researchers and clinicians interested in spoken language. Each of the 15 subcomponents of TalkBank targets a different research community or clinical interest. They are AphasiaBank for aphasia, PhonBank for child phonological development, FluencyBank for childhood disfluency, CABank for Conversation Analysis, BilingBank for bilingualism, SLABank for second language learning, CHILDES for child language development, RHDBank for right hemisphere disorder, DementiaBank for dementia, SamtaleBank for Danish, ClassBank for classroom discourse, HomeBank for daylong recordings In the home, and ASDBank for autism. Links to each of these banks can be found at http://talkbank.org.

Users of TalkBank data include researchers, students, and clinicians interested in language development or disorders within the areas covered currently by TalkBank.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Level of Curation Performed. Select all relevant types from:

D. Data-level curation – as in C above; but with additional editing of deposited data for accuracy

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Comments

TalkBank is unique in language repositories in its use of a single data format across all corpora.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Insource/Outsource Partners. If applicable, please list them.

We rely on the CMU Cloud facility for web access and storage. Material is curated on machines in our offices outside of the CMU Cloud and then our deploy system moves the data to the CMU Cloud machines. This is simply a storage and access service provided by our University. We have complete control of the four machines that are used. The one with the largest storage and memory (https://homebank.talkbank.org) has virtual machines included by the Apache httpd.conf file for each of the 12 other TalkBank repositories. We also have a machine dedicated to Git repositories and one dedicated to serving the TalkBank database. Because this is a University-internal service, there are no formal service agreements. However, a description of the service can be found at https://talkbank.org/info/CMU_Cloud.docx.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Summary of Significant Changes Since Last Application (if applicable).

We keep adding new data.

We have also built a corpus search engine and extended our analysis programs.

We have integrated our data validation and deployment system.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

Other Relevant Information.

TalkBank is a member of the CLARIN Federation.

TalkBank data has been used in over 8500 published articles.

TalkBank has received continuous funding from NIH and NSF for 36 years.

Reviewer Entry

Reviewer 1

Comments:
Accept

Reviewer 2

Comments:
Accept

ORGANIZATIONAL INFRASTRUCTURE

1. Mission/Scope

R1. The repository has an explicit mission to provide access to and preserve data in its domain.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

The TalkBank mission statement (https://talkbank.org/share/mission.html) states that:

The mission of TalkBank is to provide maximally open access to shared transcribed recordings of naturalistic spoken language interactions, often linked to media. TalkBank seeks to comply and implement both the FAIR Data Principles (Wilkinson, M. D., et al. 2016, DOI: 10.1038/sdata.2016.18) that specify that data be Findable, Accessible, Interoperable, and Reusable, as well as the TRUST Principles (Lin, Dawei et al. 2020, DOI: 10.1038/s41597-020-0486-7) which specify that the repository should commit to Transparency, Responsibility, User community, Sustainability, and Technology.

Preservation and continued access to the data is an explicit role of the repository. The CMU Library has made a commitment to long-term global web access to the transcript and media data in TalkBank, as documented at https://talkbank.org/info/CMU_Support.pdf

By fulfilling this mission, TalkBank serves to advance our understanding of the many complex features of human communication, as well as the ways in which it is learned, processed, and changed over time.

TalkBank also provides resources for researchers and clinicians seeking to understand and evaluate language disorders. These goals are supported by both the wider scientific community and funding agencies. This mission statement has been included in requests for funding across a period of 28 years and has been repeatedly supported by both NIH and NSF.

TalkBank depends on a high 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.

Further explanation of TalkBank principles is provided in (MacWhinney, 2018, DOI: 10.3758/s13428-018-1174-9)

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

2. Licenses

R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

- TalkBank data are either open access or open to researchers who are easily granted an access password. Passwords are required for access to clinical data sets, but not for the non-clinical data sets. For all data, users are expected to honor the guidelines for data use at https://talkbank.org/share/ The Ground Rules at https://talkbank.org/share/rules.html state that password-protected data cannot be posted at other sites.

- TalkBank data is available for non-commercial use through the Creative Commons BY-NC-SA 3.0 license https://creativecommons.org/licenses/by-nc-sa/3.0/

- NIH has established seven additional considerations for repositories storing human data (https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html). In accord with these principles, TalkBank tracks downloads through Google Analytics. It provides access in accord with informed consent permission, as well as identifiability and sensitivity of data. Point 7 of the TalkBank Code of Ethics at https://talkbank.org/share/ethics.html describes procedures for dealing with violations, which have been quite rare.

- All TalkBank data have received Institutional Review Board review and approval, as indicated by contribution forms at each of the 14 web sites, such as this one for CHILDES https://childes.talkbank.org/permissions/

Reviewer Entry

Reviewer 1

Comments:
The only remaining issue here are the CC-licenses for restricted files, which, as noted before, should not stand in the way of certification, so this is just for future consideration:

The CC license is not strict enough for restricted files, because it leaves the repository without a legal means to force takedown if restricted data from its holdings were to appear online: E.g., with the CC license, it would be completely legal for a publisher to publish the entirely of a restricted project online (and you wouldn't likely be able to tell how they obtained it), and you would likely have no legal means of getting them to remove it. Keeping copyright with a more restrictive license would allow for, e.g. a DCMA takedown request in the US. I realize these aren't terribly likely scenarios, but as repositories we're charged with thinking through tail risks.
This

Reviewer 2

Comments:
Accept

3. Continuity of access

R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

• TalkBank assumes responsibility for continued data preservation. There are no imposed limits on preservation, once data are
included in TalkBank. However, contributors can request removal of a corpus. Also, some corpora are contributed with an embargo period
that blocks access until research reports based on the data have been published.

• Carnegie Mellon Libraries has provided a commitment for long-term preservation of TalkBank as documented in the letter at https://talkbank.org/info/CMU_Support.pdf

• Medium-term availability of the data is guaranteed by current NIH and NSF funding that extends to four more years. Long-term preservation without new additions of data is guaranteed by CMU Libraries. Continued development of particular components of TalkBank is facilitated, but not guaranteed by commitments from members of the TalkBank Governing Board (particularly younger members) who specialize in particular types of discourse that are important to their research areas. Currently funded proposals include resources for encouraging this type of succession of TalkBank components to younger scholars. In addition, work on complete packaging of TalkBank software will make it easy to set up TalkBank components at other institutions.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

4. Confidentiality/Ethics

R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

• The repository is supported by Carnegie Mellon University, which is the relevant legal entity in contractual matters.
• We use a standard data contribution form (https://talkbank.org/share/permission.pdf).
• Data consumers are asked to follow our usage guidelines (https://talkbank.org/share/).
• All TalkBank data have received informed consent for inclusion in TalkBank and IRB review. Most data are de-identified by the contributor, avoiding disclosure risks. Data that are not de-identified are contributed with full consent. TalkBank data are in accord with GDPR requirements regarding informed consent and contributors' or participants' ability to withdraw data if needed.
• Additional conditions applying to the HomeBank unvetted audio recordings are explained at the HomeBank membership page (https://homebank.talkbank.org/membership.html) .
• If conditions would not be met, we would make the specifics of the 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.
• 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 for each repository.
• Data in the clinical databases (AphasiaBank, FluencyBank, RHDBank, and TBIBank) are password protected. About 3% of the data in other areas are in this category. This is explained in detail (https://talkbank.org/share/irb/options.html).
• Data with disclosure risk are password protected.
• Data with levels of disclosure risk beyond that of password protection are archived but not distributed.
• Files are anonymized through replacement of lastnames with the word LastName and replacement of addresses with the word Address. Audio matching these strings is replaced in a waveform editor by silence.

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

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
It is recommended that GDPR compliance is demonstrated in a more detailed way in the future.

5. Organizational infrastructure

R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

- TalkBank is hosted by Carnegie Mellon University, a highly ranked University, particularly in the areas of Computer Science and Psychology.
- TalkBank is funded by three grants from NIH and two grants from NSF. These grants expire between 2020 and 2023, but can be extended further, based on review. The FluencyBank and PhonBank projects are soon to expire, but are being submitted for renewal. In July, NSF funded a 3-year grant for core TalkBank development. The AphasiaBank project will expire next year and a renewal proposal will be submitted in November.
- We have one fulltime research assistant, three fulltime programmers, each with advanced degrees in Computer Science, and one Ph.D. (Davida Fromm) with a degree in Speech and Language Pathology.

- All personnel are able to take advanced classes at the University. We also rely on resources and collaborations from CMU's Language Technologies Institute.
- The TalkBank Governing Board (https://talkbank.org/share/governing.html) meets twice annually over Zoom to provide direction for TalkBank policies and initiatives. Additional decision-making and governance are based on an ongoing response to community input from the Google Groups lists (info-childes@googlegroups.com and others), data contributors, data users, and users of the programs. Day-to-day decisions are made by MacWhinney, Fromm, and the programmers in communication with Nan Bernstein Ratner for FluencyBank, Yvan Rose for PhonBank, Johannes Wagner for CABank, Davida Fromm and Heather Wright for AphasiaBank, Jamila Minga for RHDBank, Leanne Togher for TBIBank, Joan Kelly Hall for ClassBank, Catherine Tamsi-Lamonda for CHILDES, Anne Warlaumont and Mark VanDam for HomeBank, and Saturnino Luz and Alyssa Lanzi for DementiaBank. Decisions regarding overall TalkBank initiatives are reviewed by the Governing Board, chaired by Nan Bernstein Ratner, often done through email.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:

6. Expert guidance

R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either inhouse or external, including scientific guidance, if relevant).

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

TalkBank relies on guidance from three sources:

- Our staff, including our three professional programmers,continually interact with our colleagues in the Language Technologies Institute. Specifically these LTI faculty have been co-P.I.s on TalkBank grants: Jaime Carbonell, Ed Hovy, Eric Nyberg, Lori Levin, Alon Lavie, Maxine Eskenazi, and Florian Metze (http://www.lti.cs.cmu.edu).
- The TalkBank Governing Board as listed at https://talkbank.org/share/governing.html meets twice annually over Zoom to provide direction for TalkBank policies and initiatives.
- Additional decision-making and governance are based on an ongoing response to community input from the Google Groups lists (info-childes@googlegroups.com and others), data contributors, data users, and users of the programs.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

DIGITAL OBJECT MANAGEMENT

7. Data integrity and authenticity

R7. The repository guarantees the integrity and authenticity of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

We guarantee data integrity through a range of procedures:

- We do not enforce data fixity, because we make improvements to transcriptions based on relistening to the audio and refinement of particular codes on a lexical basis. However, we maintain copies of original contributions.
- In accord with RDSA guidelines, fixity checking for media in the CMU Campus Cloud Plus is done through the Campus Cloud Plus software.
- The completeness of the metadata is monitored by the CLARIN Centre at the Austrian Academy of Sciences that checks our CMDI metadata for completeness in accord with CLARIN standards and procedures described at https://curation.clarin.eu/.
- Changes to the transcript data and metadata are logged in GIT histories. Changes and new commits of data and media are only made by the Director. We have also now implemented version tracking and recovery using the new TalkBankDB system https://talkbank.org/DB which tags and archives each update on a file-by-file basis and which can be retrieved by the user without administrator intervention.
- Our CHAT XML Schema has been adopted by many projects as a standard for data transcription. CHAT itself includes a variety of international standards such as IPA, ISO-639, and Jeffersonian CA (https:/ca.talkbank.org/codes.html) .
- The process of data changes is communicated to users through emails to chibolts@groups.google.com. These changes never lose information, they only add it.
- We use the Scrutiny Link checker to make sure all HTML links are operative.
- PID is generated through Handle Server from which we produce OAI-PMH data.
- Corpora are registered for DOI with the assistance of the CMU Library.
- Provenance is documented through the corpus pages, such as https://childes.talkbank.org/access.
- OAI-PMH compliant CMDI metadata is created, published, and harvested through the CLARIN/CMDI system.
- Use of GIT allows us to maintain a definitive "origin" version of each file. Users may make requests to obtain versions of transcripts from a given date. It is also possible to specify the shape of the database at a given date through TalkBankDB.
- The identities of depositors are carefully checked through phone calls and emails. We have either met every one of the contributors or else spoken with them on the phone. We also check their university web sites in order to construct the various personalized web pages, such as the one at http://childes.talkbank.org/access/Biling/YipMatthews.html We stay in constant contact with all (living) contributors.
- We verify integrity through roundtrip checking of the data from XML to CHAT and then checking of the identity of the resultant roundtrip. This uses the Chatter checker available from https://talkbank.org/software/chatter.html. Corruption for language data is very different from corruption of things like spreadsheet data and the issues are quite different. SHA 256 would not be relevant to our work.

Reviewer Entry

Reviewer 1

Comments:
There's a typo: RDSA guidelines should (I assume) be "NDSA guidelines"

Reviewer 2

Comments:
Accept

8. Appraisal

R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

• Our collection development policy focuses on ingestion of transcripts that are in CHAT format along with media. The media can be in various formats (which we will convert if necessary), but the transcripts must be in CHAT format. For some important data sets, we accept untranscribed media and create the transcripts ourselves, but this is overwhelmingly the exception.
• All corpora must pass through validation by the Chatter XML validator.
• Data that do not pass Chatter are returned to the contributors, although we also often help them with the process of providing fully valid data. In some cases, the changes are minimal. Once contributors agree to needed format changes, we accept the corpus.
• All corpora must have the metadata required for the IMDI standard.
• If required metadata fields are missing, we require contributors to provide the information. However, the basic standards of CHAT and our procedure for web-page documentation make it nearly impossible for contributed corpora to have missing metadata.
• We only accept data in CHAT format as specified in the CHAT manual at https://talkbank.org/manuals/CHAT.pdf. If data are not in CHAT format, they are reformatted. If reformatting is not possible, data are not included.
• We use validation through Chatter for transcripts and our own system for metadata validation (rather than Arbil) to guarantee that data are in the correct format.
• No incorrectly formatted data can enter into the system.
• Our basic file format relies on text-only Unicode files. We expect only minor changes in this format over time. To guarantee preservation of the data in terms of transcript format, we use the Chatter program at https://talkbank.org/software/chatter to make sure that theXML version (https://talkbank.org/talkbank.xsd) of the CHAT

files can be round-tripped from CHAT to XML and back without changes. We are currently rewriting Chatter to work with a JSON, rather than XML, encoding of the database. For audio, we maintain both MP3 and WAV formats, in hope that the latter could be converted without loss to any new popular formats. If only MP3 media have been contributed, we only maintain that format. For video, we focus on making sure that everything is in .H264 format.

- 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.
- We accept any spoken language data that can be placed into one of the TalkBank repositories, as long as they match the above requirements.
- The above principles are posted at https://talkbank.org/share/preservation.html .
- Data are only removed on a contributor's or participant's request. This has only happened twice in 36 years and in each case only one file was removed. This occurred before the time of creation of PIDs. If it now becomes necessary to remove a transcript of media, it is placed in an inaccessible location.
- All corpora are described by IMDI Metadata https://en.wikipedia.org/wiki/IMDI.

Reviewer Entry

Reviewer 1

Comments:
Moving files to an inaccessible location is fine, but best practices for deaccessioning would require an explanation displayed on a "tombstone" page for the PID and an update of PID metadata (in the case of DOIs).
I'd recommend adding this to the development roadmap.

Reviewer 2

Comments:
Accept

9. Documented storage procedures

R9. The repository applies documented processes and procedures in managing archival storage of the data.

Compliance Level:

3 – The repository is in the implementation phase

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:

3 – The repository is in the implementation phase
It is recommended that, in the future, main processes and procedures, and management of changes, are documented in a more detailed way.

Response:

- Our data processing methods are described at https://talkbank.org/share/workflow.html.
- Security is maintained through SSL certificates and passwords.
- We rely on the CMU Cloud service for storage and backups. CMU Cloud maintains an ongoing image backup for 30 days. This services is described at https://talkbank.org/info/CMU_Cloud.docx.
- CMU Cloud has no drive failures. All drives are RAID drives. There is a complete local copy along with another stored at a data center off campus.
- We use CMU Cloud as well as our own backup hard drives for multiple (4) copies of all data.
- We can achieve recovery using any of our three methods: CMU Cloud, GIT, and our own hard drives.
- We believe that these three methods manage all relevant risks.
- Our backup drives contain versions from six month periods. We do not require consistency; rather, these older versions are mostly interesting as documentation for older studies that used these data.
- The definitive copy of the data is the one that sits on the servers in the CMU Cloud. Data is copied to those servers from our staging area using a deployment system that includes a wide range of validation checks. The only reason for having multiple copies of the data is to guarantee against loss. We check against corruption by running materials continually through the Chatter verifier.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

10. Preservation plan

R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept with minimal revision.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept pending minor addition, please see comment.

Response:

- The CMU University Libraries have promised long-term web-accessible access to TalkBank materials. See https://talkbank.org/info/CMU_Support.pdf.
- Our preservation level for TalkBank data involves maintaining web access to all materials.
- Because our transcript files are standard text files, we do not expect obsolescence in that regard. The greatest danger for obsolescence would be for media formats. For audio, we make sure that the .wav format is maintained and that seems likely to be capable of preservation. For video, we have everything in MP4. If that format expires, video media would have to be converted to the format that succeeds it. In the past, we have converted .avi, .mpeg, and .mov to MP4.
- When depositors contribute data to TalkBank, they maintain control in the sense that they can tell us how to reconfigure the data or even remove segments. However, they do not have the ability to directly edit files. If they need to make file-level changes, they need to tell us what to change and we will do this. This only happens on occasion.
- The repository has the rights to copy, transform and store all items.
- This procedure is specified in the contribution form at https://talkbank.org/share/permission.pdf that depositors sign.

Reviewer Entry

Reviewer 1

Comments:
The response letter links to a pretty detailed preservation statement. Please include and (very) briefly summarize the same link here.

Reviewer 2

Comments:
Please add the link to https://talkbank.org/share/preservation.html

11. Data quality

R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

• The accuracy of adherence to the CHAT transcription standards is insured by validation through the Chatter XML converter/validator.
• Chatter validates the complete XML schema which deals with data accuracy on the levels of the utterance, the word, alignments of words to parts of speech, alignment of phonemes to words, and many other features of coding of spoken language data.
• The control of metadata quality is handled by a separate XML validator that replaces the Arbil validator for CMDI.
• We verify the quality of transcription by listening to segments of incoming data. Our staff is able to check transcripts in all of the major languages in the database, except for the Slavic languages and Japanese. For Japanese, we rely on checking from Susanne Miyata, who has worked with CHAT data for three decades.
• The transcription of informal spoken language can seldom be judged as 100% accurate. There are often disagreements, particularly under conditions of noise. Some of the speakers are young children who are learning to use language. Others are people with aphasia who have various problems with language. Moreover, different researchers have different goals in the creation of transcriptions. Everyone wants to get the words right, but some care about breath groups and intonation. Others want to code errors. CHAT allows for all of these possibilities. These issues are discussed in the CHAT manual at https://talkbank.org/manuals/CHAT.pdf.
• We do not seek community input regarding data quality of individual transcripts, although we work on this with the data contributor. Also, we seek and receive continual input regarding the format of CHAT coding.
• Because TalkBank has been used in over 8500 published articles, we rely on scholar.google.com to provide links to use of the TalkBank corpora. This is done by requiring users of the data to cite certain core publications, which we can then track through Google Scholar.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

12. Workflows

R12. Archiving takes place according to defined workflows from ingest to dissemination.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

• Our workflow is described at http://talkbank.org/share/workflow.html.
• The workflow for the ingestion of data in TalkBank has the following steps: contributors read the guidelines, they send mail to macw@cmu.edu, we reply, data is transferred. This is described in greater detail at https://talkbank.org/share/contrib.html
• We then check the data using CHECK and Chatter, we create metadata files for the OLAC and CMDI/CLARIN systems, 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.
• 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.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

13. Data discovery and identification

R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

• We have recently created a complete new system for data discovery within the TalkBank repositories. It is at https://talkbank.org/DB. An early version is described in this article https://psyling.talkbank.org/years/2018/TalkBankDB.pdf, although it now has many more features and full coverage of all TalkBank data, along with working links to R/Python analysis.
• We generate metadata for OAI-PMH harvesting by the CLARIN system which then feeds data into the the VLO at (https://vlo.clarin.eu/)
• CLARIN's TLA (The Linguistic Archive) and VLO (Virtual Linguistic Observatory) provide metadata searching for TalkBank materials. About a sixth of the holdings in VLO and TLA are from TalkBank.
• Machine harvesting of the metadata is done through the the CLARIN Centre at the Austrian Academy of Sciences using the OAI-PMH protocol.
• TalkBank is included in several NIH data bank registries.
• TalkBank offers DOI citations for each corpus. These are found on the HTML pages for each corpus.
• TalkBank also provides PID's through the HandleServer system. These are used for both CLARIN and DOI. We do not update PIDs for changes in content, because these involve minor syntactic modifications. Updating at this level would be counter-productive for the field.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

14. Data reuse

R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

• When data are contributed, they must have metadata in accord with the CMDI standards.
• We continually adjust to changes in metadata formats, such as the change from IMDI to CMDI that occurred about four years ago.
• We will adjust to changes in metadata format by rewriting parts of the code we use to create metadata from information already in our transcripts and corpus documentation. Development of TalkBankDB has depended heavily on such methods.
• The meaning of CMDI metadata is documented in the CMDI documentation (see https://www.clarin.eu/content/component-metadata)

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

TECHNOLOGY

15. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept with minimal revisions.

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Response:

- For the creation and publication of our metadata, we follow OAIS and OAI-PMH standards.
- For the coding of language data, we use the CHAT format that follows standards from Linguistics, Conversation Analysis, and Speech Science, as described in the CHAT, CLAN, and MOR manuals available from https://talkbank.org
- We create CMDI metadata through special purpose code that creates the required files for online harvesting.
- We continually improve our infrastructure in terms of storage, format, coverage, and quality. We are particularly interested in the use of speech processing to improve linkage of transcripts to audio and parser technology to analyze grammatical structure.
- The software we have developed is available from our server (https://talkbank.org/software) as well as GitHub (https://github.com/talkbank).
- We are currently creating a complete inventory and documentation of the software used in the deployment and checking system. Parts of this system, such as gitlab are open-source software.
- With our database in the CMU Cloud facility, we can guarantee high-level extremely fast, around-the-clock connectivity.
- CMU Cloud guarantees recovery in 24 hours. However, we have only needed to request recovery one time and then it took two hours.
- The operating system for our CMU Cloud servers is the current version of Ubuntu Linux.

- Our plans for infrastructure development focus on the integration of the deployment system, the expansion of the TalkBankDB database system at https://talkbank.org/DB, the development of a system for collaborative commentary within the TalkBank Browser, and increased use of speech technology for diarization of transcripts with media.

Reviewer Entry

Reviewer 1

Comments:

Reviewer 2

Comments:
Accept

16. Security

R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.

Compliance Level:

4 – The guideline has been fully implemented in the repository

Reviewer Entry

Reviewer 1

Comments:
4 – The guideline has been fully implemented in the repository
Accept

Reviewer 2

Comments:
4 – The guideline has been fully implemented in the repository
Accept 4 pending minor addition

Response:

- The CMU Cloud facility guarantees swift recovery of essential serves in the event of an outage. The track record of CMU Cloud is really excellent with no outages in 3 years of service. See https://talkbank.org/info/CMU_Cloud.docx
- Apart from backup through CMU Cloud, we also maintain an offline system of backup hard drives that mirror stages of the database. In addition, GIT itself provides continual backup from the beginning of the git repository and all git repos are themselves backed up across these three systems.
- The CMU Information Security services do ongoing checking of our servers for vulnerabilities. They identified vulnerabilities in some PHP code, which we then repaired. Access to media materials in the CMU Cloud Plus server area

are totally controlled by Computing Services, making breaches even less possible.

• We rely on CMU Cloud for security and disaster planning.

• Most materials are open access. Passwords are controlled through Apache at two levels. Some corpora are available to all members of that data community. A few of the corpora in HomeBank are only open to the data contributor.

• A summary of the services provided by CMU Cloud is given at https://talkbank.org/info/CMU_Cloud.docx.

• For business continuity, we rely on the continued importance of these resources to the relevant research communities as a core way of insuring funding in the future.

Reviewer Entry

Reviewer 1

Comments:
Ideally, I'd like to see a bit more about human security here: (e.g. how exactly is access to restricted data handled -- do the right people have access, do they have the right training; are any special protections applied to restricted data (encryption at rest, etc.).

It doesn't seem like this data are high risk, but I'd flag this as another area for future development (which would likely be just as much in terms of

Reviewer 2

Comments:
In your comments you state that ‘We have met several time with the CMU Cloud supervisors and we stay in contact through email’ which is information tjat should be included in the self-assessment statement. Therefore please add that you are in regular contact with the CMU Cloud supervisors (or words to that effect).

APPLICANT FEEDBACK

Comments/feedback

These Requirements are not seen as final, and we value your input to improve the CoreTrustSeal certification procedure. Any comments on the quality of the Requirements, their relevance to your organization, or any other contribution, will be considered as part of future iterations.

Response:

I only really had problems with understanding the level of detail for procedure document for R9. Proper curation of language data into a consistent format and structure is an extremely complex process. Some repositories only need to archive .zip files with some documentation, but our task is much larger.

Reviewer Entry

Reviewer 1

Comments:

Thank you for the detailed revisions and explanations. There are a couple of minor issues remaining, but they're at the level of typos, so I now recommend certifying this application.

I've left comments encouraging future development for three other answers, but don't think they should affect certifications.

Reviewer 2

Comments:
Thank you for the improved application and the comments and explanations in the separate document.