



Assessment Information

[CoreTrustSeal Requirements 2020–2026](#)

Repository:	TalkBank
Website:	https://talkbank.org
Requirements version:	CoreTrustSeal Requirements 2023-2025

This repository is owned by:	CLARIN
Created at:	Sept. 23, 2026, 12:39 a.m.

CORE TRUSTWORTHY DATA REPOSITORIES REQUIREMENTS

Background Information

Re3data Identifier

Please fill your Re3data identifier from the website: <https://www.re3data.org/>

Response:

RRID:nif-0000-00626 RRID:SCR_003242

Repository type

Please select your repository type.

Response:

- Specialist repository

Overview

Provide a short overview of key characteristics of the repository, reflecting the repository type selected. This should include information about the scope and size of data collections, data types and formats. Further contextual information may also be added.

Response:

The TalkBank system (<http://talkbank.org>) is a specialist system for spoken language data. It is the world's largest open-access integrated repository for spoken-language data. It provides language transcripts, media, and resources to support researchers in psychology, linguistics, education, computer science, and speech pathology. There are now more than 12,000 articles based on the use of TalkBank data and programs. The system emphasizes open-science principles by making its data, programs, and tools free and easily accessible. It provides data from 47 languages across 15 content areas. These corpora have been contributed by hundreds of researchers, all using a common transcription format called CHAT. The use of this common format has allowed us to construct an innovative set of resources and tools for analyses across languages, corpora, and speakers. For a review see <https://talkbank.org/0share/CDPS.pdf>

Designated Community

A clear definition of the Designated Community demonstrates that the applicant understands the scope, knowledge base, and methodologies—including preferred software/formats—of the group(s) of users at whom the curation and preservation measures are primarily targeted. The definition should be specific so that reviewers can assess whether that community is being served in the responses to other requirements.

Response:

The repository's designated community: TalkBank provides resources for all researchers and clinicians interested in spoken language. Each of the 14 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, TBIBank for Traumatic Brain Injury, 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>. All transcript data for all banks are in the standard CHAT format at <https://talkbank.org/0info/manuals/CHAT.pdf>. As shown at <https://talkbank.org/childes/access/index.html> and elsewhere, the languages in TalkBank include: Irish, Welsh, English, Dutch, Cantonese, Mandarin, Afrikaans, Korean, Indonesian, Thai, Estonian, Hungarian, Finnish, French, German, Japanese, Catalan, Italian, Portuguese, Romanian, Danish, Swedish, Icelandic, Norwegian, Bulgarian, Croatian, Czech, Polish, Russian, Serbian, Slovenian, Spanish, Arabic, Basque, Berber, Cree, Farsi, Georgian, Greek, Nguni, Jamaican, Hebrew, Quechua, Sesotho, Tagalog, Tamil, and Turkish.

Levels of Curation

Please fill you level(s) of curation.

Response:

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

TalkBank

Levels of Curation - explanation

Please add the description for your Level(s) of Curation.

Response:

All data must conform to the transcription standard given in the CHAT manual at <https://talkbank.org/0info/manuals/CHAT.pdf>. Adherence is enforced through the use of the Chatter program at <https://talkbank.org/0info/chatter.html>. Further curation involves the use of Universal Dependencies (<https://universaldependencies.org>) through the Stanza program within the Batchalign Python script from <https://github.com/talkbank>. Batchalign also allows us to perform ASR and forced alignment on the utterance and word level as described by these articles: <https://media.talkbank.org/open/years/2023/Automation.pdf> <https://media.talkbank.org/open/years/2024/LiuMacW.pdf>. Depending on the corpus, there can be further curation for orthography. All these processes are designed to serve as filters to make sure that all data contributed to and eventually included in TalkBank conforms to our very precise standards for transcription format, data quality, and organization. For each filter, if a contribution fails to match the standard, we work with the contributor to improve the coding. Often this can involve something as simple as replacing three spaces after a colon in a speaker tier with a tab. For some other aspects of curation, we use automatic processes such as ASR and UD, as described in the articles cited.

Cooperation and outsourcing to third parties, partners and host organisations

Please describe any cooperation and outsourcing to third parties, partners and host organisations.

Response:

We do not outsource, because I would say that our use of CMU Cloud is within our university. Nonetheless, I will describe this as follows. We rely on the CMU Cloud facility for web access and storage. Material is curated on machines in our offices 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 for each of the 14 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/0info/CMU_Cloud.docx. CMU is Carnegie Mellon University in Pittsburgh.

Applicants renewing their CoreTrustSeal certification: summary of significant changes since last application.

Please fill this field when you are renewing your CoreTrustSeal Certification.

This field can be marked with not applicable (N.A.) if you are acquiring a CoreTrustSeal certificate for the first time.

Response:

With additional support from the Office of Data Science Strategy ODSS at NIH, we have made enormous progress in the last few years in terms of further implementation of FAIR and TRUST standards. 1. One crucial advance was the implementation of a new system for access control that allows differentiation between nearly fully open access and higher levels of restricted access as required by IRB policies and levels of informed consent. 2. A second major advance is the shift of our infrastructure to use of the CMU Campus Cloud facility with its strong IT protection. 3. With support from NSF we developed a system for Collaborative Commentary described at <https://talkbank.org/0info/CC> 4. We also built a new database search engine at <https://talkbank.org/0info/DB> 5. Within five of the topic areas, we developed tutorial facilities called Ground Rounds, such as <https://talkbank.org/aphasia/GR/index.html> 6. We created tutorial screencasts at <https://media.talkbank.org/open/screencasts/index.html> 7. We are in the process of transitioning our infrastructure from Java and XML to Rust and JSON/JavaScript This is just a summary of some of the major improvements.

Organisational Infrastructure

R1 Mission & Scope (R01)

R01. The repository has an explicit mission to provide access to and preserve digital objects.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

The TalkBank mission statement given in the links 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,

TalkBank

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 in the Library Support evidence link. 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 the Current Directions in Psychological Science article in the evidence link.

Links:

- [Current Directions article](#)
- [CMU Cloud](#)
- [Mission Statement](#)
- [Library Support](#)

R2 Rights Management (R02)

R02. The repository maintains all applicable rights and monitors compliance.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

- TalkBank data are either open after a user establishes an account with an email for contact and monitoring or open specifically to established researchers for the clinical data sets. Access is further controlled by a specific onboarding process for HomeBank and PsychosisBank data. The procedure for HomeBank is given in the HomeBank membership procedure evidence link. For all data, users are expected to honor the guidelines for data use as shown in the Principles evidence link and Ground Rules evidence link.
- TalkBank data is available for non-commercial use through the Creative Commons BY-NC-SA 3.0 license as given in the evidence link for Creative Commons.
- NIH has established seven additional considerations for repositories storing human data (NIH_Rules link). In accord with these principles, TalkBank tracks downloads through the new DataBase Admin system with the link given below. However, you can only log in to that link, if you have already registered. This system provides access in accord with informed consent permission, as well as identifiability and sensitivity of data. Point 7 of the TalkBank Code of Ethics in the link below describes procedures for dealing with violations, which have been quite rare. The only case in 34 years was on one by a graduate student in Japan in 1995. This case of non-compliance was handled by Japanese colleagues who made sure that the graduate student involved followed guidelines going forward.
- 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 the one for CHILDES given in the Permission Examples evidence link.

Links:

- [Permissions example](#)
- [HomeBank membership procedure](#)
- [Principles](#)
- [Ground Rules](#)
- [Creative Commons](#)
- [Ethics](#)
- [Authorization System](#)
- [NIH Rules](#)

R3 Continuity of Service (R03)

R03. The Repository has a plan to ensure ongoing access to and preservation of its data and metadata.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

TalkBank

Response:

Continuity of service involves five dimensions: (1) mission and scope, (2) business stability, (3) disaster recovery, (4) format continuity, and (5) succession planning, as described here. 1. Mission and scope. Since 1984, TalkBank has had the same basic mission of providing shared data for the study of spoken language. However, the scope of TalkBank has broadened continually across these 41 years to go from just child language to now 15 language areas. This broadening of scope supports the continuity of service because TalkBank has become a part of the infrastructure of each of these 15 research communities. 2. Business stability is provided by the integration of TalkBank into the goals and workflow of four of the institutes of the National Institutes of Health: NIDCD (communicative disorders), NICHD (child development), NIA (aging), and NIMH (mental health). This is further recognized by designation of TalkBank databases as NIH-sponsored repositories. 3. Disaster recovery is provided by storage of TalkBank data in the CMU Cloud as described in the CMU Cloud evidence link and in R14. 4. Format continuity is provided by the fact that all TalkBank transcript data are in .txt UTF8 format and all media data are in either .wav audio format or .mp4 video format. These are the most widely accepted current standards and ones with most likely longevity, as described in the data preservation evidence link. 5. Succession plans include ongoing agreements for preservation of the full TalkBank system at two universities outside the USA, agreements with two other US repositories, and commitment for preservation from the Carnegie Mellon library, as explained in the data preservation and Library Support evidence links.

Links:

- [CMU Cloud](#)
- [data preservation](#)
- [Library Support](#)

R4 Legal & Ethical (R04)

R04. The repository ensures to the extent possible that data and metadata are created, curated, preserved, accessed and used in compliance with legal and ethical norms.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

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 approved by CMU IRB and legal, as shown in the Permission Form evidence link. • Data consumers are asked to follow our usage guidelines as given in the Ground Rules evidence link. • 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 must be 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. However, if a contributor from a European institution is unable to convince their institution that data-sharing through TalkBank should be allowed under GDPR rules, then we do not include that data. European institutions vary greatly in the ways that they interpret GDPR and we just have to leave it up to each institution what decisions they wish to make. • Additional conditions applying to the HomeBank unvetted audio recordings are explained at the HomeBank membership page as given in that evidence link. • 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. • Some data in HomeBank and PsychosisBank require Data Contributor permission for access. This is explained in the Options evidence link. • 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.

Links:

- [Ground Rules](#)
- [HomeBank Membership](#)
- [Options](#)
- [Permission Form](#)

R5 Governance & Resources (R05)

R05. The repository has adequate funding and sufficient numbers of staff managed through a clear system of governance to effectively carry out the mission.

Compliance level:

TalkBank

Implemented: the requirement has been fully implemented by the repository - 1

Response:

- TalkBank is hosted by Carnegie Mellon University, a highly ranked University, particularly in the areas of Computer Science and Cognitive Psychology.
- TalkBank is funded by four grants from NIH. These grants expire between 2026 and 2028, but can be extended further, based on review.
- We have four fulltime research assistants, three fulltime programmers, each with advanced degrees in Computer Science, and one Ph.D. (Davida Fromm) co-director 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 Advisory Board (see link) 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), datacontributors, data users, and users of the programs.

Links:

- [AdvisoryBoard](#)

R6 Expertise & Guidance (R06)

R06. The repository adopts mechanisms to secure ongoing expertise, guidance and feedback-either in-house, or external.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

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. • CMU provides frequent seminars and workshops on advanced tools for programming which they attend. • These past and present LTI faculty have been co-P.I.s on TalkBank grants: Jaime Carbonell, Ed Hovy, Eric Nyberg, Lori Levin, Alon Lavie, Maxine Eskenazi, Shinji Watanabe, and Florian Metze (<http://www.lti.cs.cmu.edu>). • The TalkBank Advisory Board as listed at (see link) meets twice annually over Zoom to provide direction for TalkBank policies and initiatives. • Our colleague Houjun Liu at Stanford is a member of the Stanza group that works on our taggers for morphosyntax. Houjun has also developed an advanced Python script called Batchalign at GitHub to apply ASR and Universal Dependencies tools to TalkBank data. • All our materials are open sourced at <https://github.com/talkbank> and we are major contributors to the HuggingFace diarization group. • 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.

Links:

- [GitHub](#)
- [Language Technologies Institute](#)
- [Advisory Board](#)

Digital Object Management

R7 Provenance and authenticity (R07)

R07. The repository guarantees the authenticity of the digital objects and provides provenance information.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

We guarantee provenance and authenticity through a range of procedures: • We do not enforce data fixity, except for media, 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 both in Git repositories and through ongoing backup. • In accord with RDSA guidelines, fixity checking for media in the CMU Campus Cloud Plus is done through the Campus Cloud Plus software. • 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 in consultation with the Data Contributor(s). 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 implement version

TalkBank

tracking and recovery using the TalkBankDB system (see link) 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 (Conversation Analysis) (see link) . • We use a link checker inside poetry to make sure all HTML links are operative. • PID is generated through Handle Server • Corpora are registered for DOI with the assistance of the CMU Library. • Provenance is documented through the corpus pages, such as the one in the Example Page evidence link. • 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 the Example Page evidence link. 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 the Chatter Page and Chatter validator evidence links. Corruption for language data is very different from corruption of things like spreadsheet data and the issues are quite different, and SHA 256 would not be relevant to our work.

Links:

- [Chatter validator](#)
- [TalkBankDB](#)
- [CA Codes](#)
- [Chatter Page](#)

R8 Deposit & Appraisal (R08)

R08. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for users.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

• Our collection development policy focuses on ingestion of transcripts that are in CHAT format (see CHAT manual link) along with media. The incoming media can be in various formats (which we will convert if necessary), but the transcripts must be in CHAT format. For important data sets, we accept untranscribed media and create the transcripts ourselves using ASR. We then send the resultant transcripts back to the Data Contributor for checking. • All corpora must pass through validation by the Chatter XML validator (see Chatter Page link). • 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 DOI and DataCite. • 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 (see . If data are not in CHAT format, they are reformatted. If reformatting is not possible, data are not included. • 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 to make sure that 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 store either WAV format or MP3 format, depending on what was contributed. However, for web delivery through the TalkBank Browser we use MP3. For video, we focus on making sure that everything is in .H264 (.mp4) 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. • 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 or media, it is placed in an inaccessible location.

Links:

- [Chatter Page](#)
- [CHAT Manual](#)

R9 Preservation plan (R09)

R09. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

TalkBank

Response:

• Our plans for long-term data preservation are in Data Migration and Compatibility section in the Preservation evidence link. Here is a copy of that section: Our basic file format relies on text-only UTF8 files. We expect only minor changes in this format over time. However, the CHAT coding system occasionally 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. For audio, we maintain both MP3 and WAV formats, assuming that the lossless WAV format could be converted to any new popular formats. However, some corpora come with only MP3 files. 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. 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.

Links:

- [CMU Cloud](#)
- [Preservation](#)

R10 Quality Assurance (R10)

R10. The repository addresses technical quality and standards compliance, and ensures that sufficient information is available for end users to make quality-related evaluations.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

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, comment scoping, and many other features of coding of spoken language data. • The control of metadata quality is handled by the Chatter 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 the major languages in the database, but we can also rely on colleagues to check others. • Because the majority of the transcripts are linked to the media on the utterance and often word level, and because they can easily be played back in the TalkBank Browser, the accuracy of the transcription can be directly verified by the Data Consumer. • 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. • 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 12,000 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.

Links:

- [Chatter validator](#)
- [CHAT Manual](#)

R11 Workflows (R11)

R11. Digital object management takes place according to defined workflows from deposit to access.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

• Our workflow is described in the Workflow link. h • 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, and data is transferred. This is described in greater detail in the Contributing link. • 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.

TalkBank

Links:

- [Workflow](#)
- [Contributing](#)

R12 Discovery and Identification (R12)

R12. The repository enables users to discover the digital objects and refer to them in a persistent way through proper citation.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

• We have recently created a complete new system for data discovery within the TalkBank repositories. It is at <https://talkbank.org/info/DB> and it has an API for R/Python analysis. However, because that system was receiving such heavy usage, we had to do a total rewrite that will download jobs to user computers, rather than performing all the computation on the server. We are putting the older, less capable version of the system back on the web by Dec 1, but the full new version should be available by January 15 • Users can locate TalkBank corpora through DataCite or a Google Search. For many topics, TalkBank corpora are the first hit. • From the main web site at <https://talkbank.org> there are links to each of the 15 databases. At each site there are access tables that list the available corpora. The evidence contains the URL for one such table. • 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 DOI. We do not update PIDs for changes in content, because these involve minor syntactic modifications.

Links:

- [TalkBankDB](#)

R13 Reuse (R13)

R13. The repository enables reuse of the digital objects over time, ensuring that appropriate information is available to support understanding and use.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

• With password permission, transcripts and media can be downloaded for further analysis using the CLAN programs (see link) and other programs. • The TalkBankDB system provides APIs for data analysis through R and Python. • When data are contributed, they must have metadata provided at the file, corpus, and collection levels. • At the file level, each transcript has 12 fields of required information about participants, languages, media format, and date of recording. This information is in the first lines of the file and the CLAN programs are tuned to working on this metadata. • At the corpus there is a Ometadata.cdc file for information pertaining to all transcripts in the corpus, including contributor, language, type of study, country, year of contribution, and DOI. • The corpus access page provides the fullest expository description of each corpus and this goes way beyond simple metadata to include information on tests, goals of the study, limitations, publications and much other information. The Example Page evidence link illustrates this. • The HTML pages also include some standard links that can be located through a search.

Links:

- [Example Page](#)
- [TalkBankDB](#)
- [CLAN Manual](#)

Information Technology & Security

R14 Storage & Integrity (R14)

R14. The repository applies documented processes to ensure data and metadata storage and integrity.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

TalkBank

Response:

• To ensure data integrity, we use the CHAT format that follows standards from Linguistics, Conversation Analysis, and Speech Science, as described in the CHAT and CLAN manuals (see links). • We continually improve our infrastructure in terms of storage, format, coverage, and quality. We done a lot to apply advances in ASR and NLP to our data, as described in the Automation and Morphosyntactic papers in the evidence links. • The software we have developed is available from our server through the Software evidence link as well as source code in GitHub at the GitHub evidence link. • With our database in the CMU Cloud facility (see link), 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 (see link), the further development of the Collaborative Commentary system (see link), and increased use of speech technology for diarization of transcripts with media. • We use GIT logs and facilities, along with our validator, link checker, and transcript-media co-checker, to check for ongoing corpus integrity and correctness and to allow for version rollback if problems are found.

Links:

- [CMU Cloud](#)
- [Chatter validator](#)
- [CLAN Manual](#)
- [CHAT Manual](#)
- [GitHub](#)
- [TalkBankDB](#)
- [Collaborative Commentary](#)
- [Automation](#)
- [Morphosyntactic](#)

R15 Technical Infrastructure (R15)

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

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

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 8 years of service. See https://talkbank.org/0info/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 once 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. • Password access (for currently 43,376 users) is controlled by the powerful new Admin Console (see link), which can only be accessed by system administrators. • A summary of the services provided by CMU Cloud is given at https://talkbank.org/0info/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. Given the increasing importance of TalkBank data for new AI projects using spoken language, financial support is not a problem.

Links:

- [CMU Cloud](#)
- [Admin Console](#)

R16 Security (R16)

R16. The repository protects the facility and its data, metadata, products, services, and users.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

TalkBank

- For information security, the CMU Information Security services does ongoing checking of our 8 local (desktop) machines for vulnerabilities using Falcon CrowdStrike and other monitoring facilities. None of these machines are open to the public. However, we have one MacStudio that is serving our new WebService for ASR and that is open to the public with a password. It was compromised in a recent widely-publicized vulnerability which was detected by the CMU monitoring system within hours and we shut down the machine, rebuilt it, and instituted a new system for our own quick response to vulnerabilities as reported from GitHub. We should note that this one public machine has no local data and is only running a web service.

Access to our media materials in the CMU Cloud Plus server area are controlled completely by CMU Cloud Computing Services, making breaches less possible. All of our transcript, documentation, and HTML data are continually committed to GitHub repositories, and backed up through full mirror backups each month and stored for the last 22 years.

- Most materials are open access, but require registration as tracked by the Admin Console (see link) which implements any required further access restrictions.
- A summary of the services provided by CMU Cloud is given at that evidence link.

Links:

- [CMU Information Security Office](#)
- [Admin Console](#)

Applicant feedback

R17 Applicant Feedback

We welcome feedback on the CoreTrustSeal Requirements and the Certification procedure.

Compliance level:

Implemented: the requirement has been fully implemented by the repository - 1

Response:

This previous revision was difficult. Managing the evidence links and dealing with incorrect markings of "accepted" status was the most difficult part. Fortunately, after a break of about 6 months, Thomas was able to fix the two worst problems (old links and incorrect labeling of outcomes) and I could then quickly make progress. In the meantime, one link changed for the CDPS article because we had to deal with the size limits of Git repos. I fixed that and I also moved the link to the NIH_Rules to evidence and out of the text. Hopefully things are okay now.

There is one ongoing problem I might mention which relates to the fact that with a complex website like TalkBank things are continually being contributed and we add new software functionality. As a result, URLs often have to change and that messes up the CTS review process. I don't know how to fix this general issue.

Links: