



## **Implementation of the CoreTrustSeal**

The CoreTrustSeal board hereby confirms that the Trusted Digital repository TalkBank complies with the guidelines version 2017-2019 set by the CoreTrustSeal Board.

The afore-mentioned repository has therefore acquired the CoreTrustSeal on September 15, 2017.

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

Yours sincerely,

The CoreTrustSeal Board

## Assessment Information

Guidelines Version:	2017-2019   November 10, 2016
Guidelines Information Booklet:	<a href="#">CTS Requirements 2017-2019</a>
All Guidelines Documentation:	<a href="#">Documentation</a>
Repository:	TalkBank
Seal Acquiry Date:	Sep. 15, 2017
For the latest version of the awarded CoreTrustSeal forhis repository:	<a href="https://www.coretrustseal.org/why-certification/certified-repositories/">https://www.coretrustseal.org/why-certification/certified-repositories/</a>
Previously Acquired Seals:	Seal date: April 8, 2014 Guidelines version: 2014-2017   July 19, 2013
This repository is owned by:	<b>Carnegie Mellon University</b> 254M Baker Hall CMU - Psychology 5000 Forbes Avenue 15213 Pittsburgh PA USA  T +1 412 268-3793 E <a href="mailto:macw@cmu.edu">macw@cmu.edu</a> W <a href="http://talkbank.org/">http://talkbank.org/</a>

# Assessment

## 0. Context

### Applicant Entry

#### *Self-assessment statement:*

Repository type: Subject-based repository for spoken language.

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, LangBank for classical languages, SamtaleBank for Danish, ClassBank for classroom discourse, TutorBank for tutors, HomeBank for daylong recordings in the home, and ASDBank for autism. Links to each of these banks can be found at <http://talkbank.org>.

Level of curation performed: Level D

All transcript data is all converted into the CHAT format. <http://talkbank.org/manuals/CHAT.pdf>. In addition, some data is checked against linked audio data for accuracy.

Outsource partners: None

Other relevant information:

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 (<http://talkbank.org/styles/map.png>) shows. TalkBank data is mirrored by the NL-CLARIN Center at the MPI in Nijmegen that has the Data Seal of Approval. The only outsourcing we do is for cloud backup through backblaze.com. 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 <http://scholar.google.com> shows that there are 7450 published articles

based on use of the TalkBank databases. TalkBank websites have received over 6 million hits. Current NIH support involves four major ongoing five-year grants for child language (<http://childes.talkbank.org>), aphasia (<http://aphasia.talkbank.org>), fluency (<http://fluency.talkbank.org>), and phonology (<http://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 (<http://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.

All URLs were visited on March 12, 2017.

## **Reviewer Entry**

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 1. Mission/Scope

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

## Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

The TalkBank mission statement (<http://talkbank.org/share/mission.html>) states that:

- The mission of TalkBank is to provide access to shared transcribed recordings of naturalistic spoken language interactions. 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 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.

URLs were visited on March 12, 2017.

CoreTrustSeal Board

W [www.coretrustseal.org](http://www.coretrustseal.org)

E [info@coretrustseal.org](mailto:info@coretrustseal.org)

## **Reviewer Entry**

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 2. Licenses

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

- TalkBank data is licensed by Creative Commons CC BY-NC-SA 3.0, as indicated in the link at the bottom of the TalkBank [homepage](#).
- Password access to clinical corpora is only given to fulltime faculty or clinicians with SLP (Speech and Language Pathology) certification from ASHA (the American Speech and Hearing Association).
- Other corpora are open access.
- Members agree to the ground rules (<http://talkbank.org/share/>).

URLs were visited on March 12, 2017.

### Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

CoreTrustSeal Board

W [www.coretrustseal.org](http://www.coretrustseal.org)

E [info@coretrustseal.org](mailto:info@coretrustseal.org)

### 3. Continuity of access

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

#### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

- Mid-term preservation of TalkBank assets is guaranteed by the fact that TalkBank projects are currently funded until 2011.
- Carnegie Mellon University Libraries which has signed an agreement (<http://talkbank.org/share/CMU-CLARIN.pdf>) to be a member of the CLARIN Federation and to continue support for TalkBank assets. The receipt of the signed contract is still pending, awaiting approval from the CLARIN General Assembly meeting in September.
- Carnegie Mellon Libraries, has established a system for long-term data preservation (LTDP) of resources created at the University, including TalkBank. We are working on the process of including all of TalkBank in this system.
- In addition to preservation at Carnegie Mellon, all TalkBank materials are included in the TLA (The Language Archive) archive (<http://tla.mpi.nl>) as a part of the CLARIN system.
- When the current director, Brian MacWhinney, retires in 2024, the current Director of the AphasiaBank Project, Davida Fromm, will assume the role of Director of that project. Yvan Rose of Memorial University Newfoundland will assume directorship of the PhonBank component. Johannes Wagner of Southern Denmark University will assume directorship of CABank. Nan Bernstein Ratner will assume directorship of FluencyBank. Fromm, Rose, Wagner, and Ratner will continue coordination of efforts through Fromm at CMU.



URLs were visited on March 12, 2017.

### **Reviewer Entry**

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 4. Confidentiality/Ethics

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

- The repository is supported by Carnegie Mellon University, which is the relevant legal entity in contractual matters.
- We use a standard data contribution form (<http://talkbank.org/share/release.pdf>).
- Data consumers are asked to follow our usage guidelines (<http://talkbank.org/share/>).
- Additional conditions applying to the HomeBank unvetted audio recordings are explained at the HomeBank membership page (<http://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.
- All data in AphasiaBank are password protected. About 3% of the data in other areas are in this category. This is explained in detail (<http://talkbank.org/share/irb/options.html>).

**CoreTrustSeal Board**

W [www.coretrustseal.org](http://www.coretrustseal.org)

E [info@coretrustseal.org](mailto:info@coretrustseal.org)

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

URLs were visited on March 12, 2017.

### **Reviewer Entry**

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 5. Organizational infrastructure

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

TalkBank is located at 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, totalling \$5,000,000. These grants expire between 2020 and 2023, but can be extended further, based on review. Reviews of proposals at NIH have never been rejected. Reviews at NSF are successful 50% of the time. We have three fulltime research assistants, three fulltime programmers, each with advanced degrees in Computer Science, and two clinical researchers, one with Ph.D. and one with SLP (M.S and certification 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.

### Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 6. Expert guidance

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

TalkBank relies on guidance from three sources:

1. 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 Metzger (<http://www.lti.cs.cmu.edu>).

2. Each of the funded TalkBank projects has an Advisory Board from which we solicit advice and input. Advisory Boards have between 10 and 15 members, selected from fulltime academics working in the relevant subareas with an interest in data-sharing. We do not rely on these people for programming expertise, but rather for subject matter expertise and direction. HomeBank relies on the already established DARCLE group. AphasiaBank (NIH grant) has had meetings or workshops five times and uses group calls in other years. FluencyBank (NSF and NIH grants) is currently organizing its Advisory Board and will operate through meetings at ASHA (American Speech and Hearing Association) and web conferencing in small groups. PhonBank and CHILDES rely on yearly meetings either at the Boston University Child Language Conference or the International Association for the Study of Child Language. These yearly sessions include updates on the project and requests for new developments. We are also in constant contact with all these people through our four Google Groups mailing lists and the DARCLE list.

3. We rely heavily on user input from our four mailing lists at [groups.google.com](http://groups.google.com)

4. In response to the reviewer comment on this issue, I need to clarify that nearly none of the experts involved in TalkBank are members of the CLARIN community. CLARIN is primarily a European entity at this point and TalkBank is located in the United States. Of course, we rely on CLARIN for ideas whenever possible. For example, we use CLARIN formats for CMDI and the new MTAS database search engine facility. However, we are relying on dozens of experts in the 12 different "banks" involved in TalkBank for subject matter and computational expertise. Three of the banks have formal Advisory Boards (AphasiaBank, PhonBank, and FluencyBank), but others rely on more distributed input for things like data format, coverage, IRB issues, rights, etc. At CMU, we rely on experts in the Computer Science department and the Language Technologies Institute which have advanced groups in areas such as speech technology, NLP, and so on.

**CoreTrustSeal Board**

W [www.coretrustseal.org](http://www.coretrustseal.org)

E [info@coretrustseal.org](mailto:info@coretrustseal.org)

## **Reviewer Entry**

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 7. Data integrity and authenticity

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

- To ensure non-corruption, we use the Chatter XML validator. Chatter serves not only as a validator, but can also produce a roundtrip from the readable CHAT format, to XML, back to CHAT and then the final comparison of the CHAT versions to guarantee that no information has been lost.
- 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.
- The completeness of the metadata is being monitored now by a new system at the Austrian Academy of Sciences that checks our CMDI metadata for completeness. We completed the first of these completeness checks in January 2017.
- Changes to the transcript data and metadata are logged in GIT histories, as well as through backblaze.com backups.
- 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 (<http://talkbank.org/CABank/codes.html>).
- The process of data changes is communicated to users through emails to [chibolts@groups.google.com](mailto:chibolts@groups.google.com). These changes never lose information, they only add it.

- Provenance is documented through the corpus pages (<http://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.
- 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. 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.

URLs were visited on March 12, 2017.

Response to reviewer question: TalkBank has 5TB of media (audio and video data) and 1.5 TB of transcript data.

## **Reviewer Entry**

*Accept or send back to applicant for modification:*

Accept

*Comments:*



## 8. Appraisal

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

- 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.
- 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 CMDI 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.
- We use validation through Chatter for transcripts and Arbil (<https://tla.mpi.nl/tools/tla-tools/arbil/>) for metadata 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 <http://talkbank.org/software/chatter> to make sure that the XML version (<http://talkbank.org/talkbank.xsd>) of the CHAT files can be roundtripped from CHAT to XML and back without changes. 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 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.
- These principles are posted here (<http://talkbank.org/share/preservation.html>) .

URLs were visited on March 12, 2017.

## **Reviewer Entry**

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 9. Documented storage procedures

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

- Our data processing methods are described here (<http://talkbank.org/share/workflow.html>).
- Security is maintained through ssl certificates and passwords.
- Because the data are open access, we are not concerned with the possibility that people may access the data.
- We currently rely on backblaze.com for backups. Backblaze stores a complete set of all versions of all files without expiration.
- We are in the process now of shifting data storage to the CMU Cloud, (<http://www.cmu.edu/computing/repair/colocation/campus-cloud>). Our server at <http://sla2.talkbank.org> is the first instance of this and two more will be added this week. CMU Cloud maintains an ongoing image backup for 30 days.
- We use backblaze.com 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, backblaze.com, and our own hard drives.

**CoreTrustSeal Board**

W [www.coretrustseal.org](http://www.coretrustseal.org)

E [info@coretrustseal.org](mailto:info@coretrustseal.org)

- 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 now copied to those servers from our staging area using rsync. 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.
- We have only experienced one drive failure once, because our drives are not online. They are mirror copies that sit on the shelf. Hard drive copies do not deteriorate. They only experience catastrophic failure. Backblaze copies do not deteriorate.

URLs were visited on March 12, 2017.

## **Reviewer Entry**

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 10. Preservation plan

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

- The CMU University Libraries have currently signed on as the supporting agency for our membership in CLARIN. It is also their policy to preserve materials in the long-term. However, the details of how this will work out for TalkBank are still being worked out.
- Our preservation level for TalkBank data involves maintaining web access to all materials.
- When depositors contribute data to TalkBank, they no longer maintain control of the data.
- The repository has the rights to copy, transform and store all items.
- This procedure is specified in the contribution form (<http://talkbank.org/share/release.pdf>) that depositors sign.

URLs were visited on March 12, 2017.

Response to reviewer question on point #3: Depositors 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.

### **Reviewer Entry**

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 11. Data quality

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

- Data quality is insured on the syntactic level by validation through the Chatter XML converter/validator. Metadata validity is insured by processing through CLARIN.
- These processes are all automated.
- We do not seek community input regarding data quality. However, we do seek and receive continual input regarding the format of CHAT coding.
- Because TalkBank has been used in over 8000 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.

Response to reviewer query: The Chatter XML validate goes far beyond control of the quality of metadata. In fact the control of metadata quality is handled by a separate XML validator that replaces the (rather buggy) Arbil validator for CMDI. 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. We verify the quality of transcription by listening to segments of incoming data. We do not accept contributions that are not completely transcribed.

### Reviewer Entry

CoreTrustSeal Board

W [www.coretrustseal.org](http://www.coretrustseal.org)

E [info@coretrustseal.org](mailto:info@coretrustseal.org)

*Accept or send back to applicant for modification:*

Accept

*Comments:*



## 12. Workflows

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

Our workflow is described here (<http://talkbank.org/share/workflow.html>).

URLs were visited on March 12, 2017.

### Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

### 13. Data discovery and identification

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

#### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

- Corpora for particular projects can be located through a hierarchical series of HTML access tables for such as as those at (<http://talkbank.org/access>), (<http://childes.talkbank.org/access>), and (<http://homebank.talkbank.org/access>)
- Each of these access tables provides for each corpus: an HTML page documenting the nature of the corpus, a downloadable .zip file with the transcripts, and a web page for downloading of the media.
- The repository can be searched directly using file by file or through the search commands built into the TalkBank browser (<http://talkbank.org/browser>) window at the bottom left.
- Using a PEPPER (<http://corpus-tools.org/pepper>) input module, we have included TalkBank corpora into the ANNIS system <http://corpus-tools.org/annis> for corpus analytic searching.
- 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 holding in VLO and TLA are from TalkBank.
- Machine harvesting of the metadata is done through the CLARIN ARBIL/CMDI OAI-PMH system.

CoreTrustSeal Board

W [www.coretrustseal.org](http://www.coretrustseal.org)

E [info@coretrustseal.org](mailto:info@coretrustseal.org)

- 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.

URLs were visited on March 12, 2017.

### **Reviewer Entry**

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 14. Data reuse

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

- 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.
- The meaning of CMDI metadata is documented in the CMDI documentation (see <https://www.clarin.eu/content/component-metadata> )

### Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 15. Technical infrastructure

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

- For the creation and publication of our metadata, we follow OAIS and OAI-PMH standards.
- We implemented this 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.
- Our software is available from our server (<http://talkbank.org/software>) as well as GitHub (<http://github.com/talkbank>).
- We are not relying on any community-supported software.
- As we move our servers to the CMU Cloud facility, we can better guarantee high-level around-the-clock connectivity.

### Reviewer Entry

CoreTrustSeal Board

W [www.coretrustseal.org](http://www.coretrustseal.org)

E [info@coretrustseal.org](mailto:info@coretrustseal.org)

*Accept or send back to applicant for modification:*

Accept

*Comments:*

## 16. Security

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### Applicant Entry

*Statement of Compliance:*

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

*Self-assessment statement:*

- The current movement of our servers to the CMU Cloud facility will guarantee swift recovery of essential services in the event of an outage. The track record of CMU Cloud is really excellent with no outages in 3 years of service.
- We will increasingly rely on CMU Cloud for security and disaster planning.
- A summary of the services provided by CMU Cloud is given at <http://talkbank.org/share/CMUCloud.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.

Response to reviewer query: Apart from backup through CMU Cloud and Backblaze, 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.

### Reviewer Entry

*Accept or send back to applicant for modification:*

Accept

CoreTrustSeal Board

W [www.coretrustseal.org](http://www.coretrustseal.org)

E [info@coretrustseal.org](mailto:info@coretrustseal.org)

*Comments:*



## **17. Comments/feedback**

*Minimum Required Statement of Compliance:*

0. N/A: Not Applicable.

### **Applicant Entry**

*Statement of Compliance:*

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

*Self-assessment statement:*

The hyperlink facility in the editor was not functioning properly, so I have added the links directly to the text. Doing this was not easy. There is no documentation about how to do this. I remember having this problem last time. It is remarkable that it has not been fixed.

### **Reviewer Entry**

*Accept or send back to applicant for modification:*

Accept

*Comments:*