Data Curation Profile – Linguistics

Profile Author	Kornelia Tancheva
Author's Institution	Cornell University
Contact	Kornelia Tancheva; kt18@cornell.edu
Researcher(s) Interviewed	Withheld.
Researcher's Institution	Cornell University
Date of Creation	3/31/2012
Date of Last Update	
Version of the Tool	1.0 / modified
Version of the Content	
Discipline / Sub-Discipline	Linguistics
Sources of Information	Interview conducted on March 23, 2012. A worksheet completed by the interviewer as a part of the interviews.
Notes	
URL	http://www.datacurationprofiles.org http://hdl.handle.net/1813/29064
Licensing	This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States License.

Section 1 - Brief summary of data curation needs

Currently the researcher is segmenting and transcribing her Cheyenne/English language audio files herself. The process is far from ideal; she is applying for a grant that would allow her to hire a lab technician, ideally with linguistic background or training that would help her do that; even if it is only for the parts that are in English. Once the files have been transcribed; the data needs to be cleaned up and normalized; and metadata applied, after which it should be searchable, so that the textual files can be searched for specific linguistic features and then call up the audio files. Ultimately the data will be ingested in a publically accessible searchable db, that allows for the download of segments. The db is searchable in the original language, in English and by the morphological gloss (which is the closest notion to a data dictionary in linguistics.) The data needs to be backed up and preserved indefinitely.

Section 2 - Overview of the research

The research project is to understand various linguistic aspects of the Cheyenne language through recordings of short stories or examples given to particular questions; the goal is to learn about the linguistic aspects of the language with a focus on semantics, pragmatics, and some syntactic issues; the data, however can be used by other linguists and the researcher has collaborated on phonetic aspects with other researchers.

http://dx.doi.org/10.5703/1288284315007

2.1 - Research area focus
Cheyenne language semantics, pragmatics and syntax.

2.2 - Intended audiences
Other linguists and potential language learners.

2.3 - Funding sources
Small grants from Rutgers and Cornell, as well as the Endangered Languages Fund; and a Philips grant from the American Philosophical Society.

Section 3 - Data kinds and stages

3.1 - Data narrative
There are two kinds of data that the researcher describes: the actual data she is working with and the ideal-state data that she would like to work with.

The actual raw data are 75 audio files; and about 20 excerpts. New files are being added each year. The ideal data include: the 75 raw files; that are normalized and cleaned up; then transcribed using a transcription software, ideally ELan, and metadata created using morphological glosses; so that the textual files can be searched for either the native language; the English; or the gloss. The audio and the transcription are synchronized. Ideally, the data at this stage is publicly available in a searchable database.

The interview focused on the ideal state of the data.

3.2 – The data table

Data Stage	Output	# of Files / Typical Size	Format	Other / Notes
Primary Data				
Raw	audio	75/2-3 gigs each	wav	
Processed	audio	75	wav	Cleaned up, normalized data
Analyzed	Audio+text	Unknown at this point	Wav + word	Transcribed normalized data; synchronized audio and text; metadata applied, including morphological glosses; searchable
Finalized	db	unknown	php	Searchable and downloadable in segments
Ancillary Data				
Ancillary Data #1	Morphological glosses			Standard online resources
Ancillary Data #2				

Note: The data specifically designated by the scientist to make publicly available are indicated by the rows shaded in gray. Empty cells represent cases in which information was not collected or the scientist could not provide a response.

3.3. - Target data for sharing

The researchers is willing to share (and has shared segments of the raw data with immediate collaborators); is willing to share the cleaned up, normalized and transcribed data with immediate collaborators and other researchers at the institution and the filed; the publicly available data will be in its final stage, i.e. in a linguistic database that she has created.

3.4 - Value of the data

The data could be of use to other linguists, as well as to language learners.

3.5 - Contextual narrative

Sharing of research data is still not the norm in linguistics, even though there is a move to get some journals to include it in supplementary form online; but this has not happened yet.

Section 4 - Intellectual property context and information

The data are not under copyright restrictions.

4.1 - Data owner(s)

The data set belongs to the researcher since no funding agency imposes any restrictions. The question is not particularly relevant to the data since the language speakers can also be viewed as the "owners" of the data. The interviewees have signed release forms, though.

4.2 - Stakeholders

None of the funding agencies impose any restrictions even though one of them, the Endangered Language Fund has offered to preserve copies of the raw data.

4.3 - Terms of use (conditions for access and (re)use)

Freely available.

4.4 - Attribution

The researcher would like the database to be cited when the data are used.

Section 5 - Organization and description of data (incl. metadata)

One of the biggest problems identified by the researcher is the lack of standardization in linguistic metadata, which may be partially due to the very narrow and specific nature of each researcher's project. Possible crosswalks envisioned.

5.1 - Overview of data organization and description (metadata)

Currently the data is in audio files (wav format), as well as segments transcribed in a word processing program (Latex).

In its ideal state, it should be described by morphological glosses, which a standard linguistic practice.

The data should be available in multiple formats (i.e. audio and text).

5.2 - Formal standards used

Currently not used; ideally morphological glosses should be used to describe the data.

5.3 - Locally developed standards

N/A

5.4 - Crosswalks

N/A currently, but if the data is part of a large language repository they will be necessary.

5.5 - Documentation of data organization/description

Not kept.

Section 6 - Ingest / Transfer

Data clean-up and normalization; transcription using ELan to synchronize the audio and the transcription; metadata.

Section 7 – Sharing & Access

In the final stage, the data is freely accessible online for any researcher or language learner.

7.1 - Willingness / Motivations to share

The researcher is willing to share the data, although she cautions that the raw files are very long and “dense” to be of use to other researchers.

7.2 - Embargo

No.

7.3 - Access control

No restrictions required, except for password creation and use to track the number of visitors and users of the data.

7.4 Secondary (Mirror) site

Initially not considered important; after prompting, researcher agrees it could be important.

Section 8 - Discovery

Should be discoverable through Google, the researcher’s web page, the department’s web page; and on the language page.

Section 9 - Tools

Portable audio recorder; computer; listening software (Audacity); transcription software (ELan); web browser to use the db

Section 10 – Linking / Interoperability

Low priority but the reason may be the lack of a comprehensive language db; or the specificity of each researcher’s project; as well as the possible lack of standardization.

Section 11 - Measuring Impact

No particular impact is envisioned—whether other researchers use the data, or how many, or what their affiliation is, will not change the researcher’s interest in the data.

11.1 - Usage statistics & other identified metrics

Number of visitors and users of the db.

11.2 - Gathering information about users

Mildly interested in their affiliation and interests (whether they are researchers or language learners)

Section 12 – Data Management

Stored on computer hard drive and external discs or hard drive. No particular security desired.

12.1 - Security / Back-ups

Computer files protected by password on the computer. Computer files backed up every day. External discs not backed up.

12.2 - Secondary storage sites

N/A

12.3 - Version control

The data do not have different versions; new data is being added to the data set but no data are being changed.

Section 13 - Preservation

The data is of potential use for the foreseeable future since it can be used for a variety of purposes and hence should be preserved indefinitely.

13.1 - Duration of preservation

Indefinitely.

13.2 - Data provenance

Not discussed.

13.3 - Data audits

Raw data integrity is crucial and hence it should be periodically audited.

13.4 - Format migration

Not discussed.

Section 14 – Personnel

This section is to be used to document roles and responsibilities of the people involved in the stewardship of this data. For this particular profile, information was gathered as a part of a study directed by human subject guidelines and therefore we are not able to populate the fields in this section.