Sarah Moeller  sarah_moeller@gial.edu
Graduate Institute of Applied Linguistics
4th International Conference on Language Documentation and Conservation, February 26-March 1, 2015, University of Hawaii at Manoa

SAYMORE: A LANGUAGE DOCUMENTATION TOOL
THE SITUATION

- Language Documentation methods and skills are not difficult to learn
- Citizen scientists could make a huge contribution to language documentation
- However, the people engaged in language documentation are largely the same people engaged in language description
THE PROBLEM

- Citizen scientists may determine that the current bar for entry is too high

- Not because the methods and skills are hard to acquire, but because of “nitty-gritty” technical details.
  - Specifically: Confusion about metadata, complexities of file management, an intimidating archiving process
Confusion about metadata:
What is metadata?
What to gather? How do I know if I have enough? How do I know if I have too much?
What to label it, e.g. “contributor” or “participant” or “speaker”?
OLAC or IMDI metadata standards – active knowledge means knowing how to use it correctly, as opposed to passive knowledge which means knowing what it is and which is which.
XML is a coding format that is both human- and machine-readable.
How to deal with access management?
Storing informed consent
Access protocol – what levels of access restriction? What is the archive’s protocol?
Complexities of File Management

- Constantly growing number of files
- Need to keep related files bundled together
- Human error (e.g. typos in filenames) leads to confusion and broken links
FILE MANAGEMENT: MANUAL METHODS

...\ALDP\ can contain subfolders like:

Admin\ (Reports, Research Permissions, Lists of files and sessions, Invoices...)
Contains folders such as:
VWS, FUNAI, Internal, Data, Financial
Papers+Results\ (scientific products)
Contains folders such as:
2008-06-DOBES-filemanagement
MPI-Corpus\ (the final data tree for uploading)
Media\ (all audio and video files, images...)
Data\ (files related to annotation and processing)


This method uses the folder directory on your laptop (e.g. My Documents). A great system. However, who doesn’t sometimes drag and drop files to the wrong folder? Or gives files in different folders the same name and then not know which one goes where? Or skips a number while naming files sequentially, meaning that that file and all subsequent ones must be renamed?
Most archives make it clear on their website that they are primarily interested in receiving and preserving otherwise endangered data in any format. The picture is not all bad, but it could be better for depositors. And the archivists’ work could be made easier.

- Finding and identifying the most appropriate archive for one’s project
- Preparing corpus for depositing
- Satisfying the archive’s preferences
SayMore is an SIL software, still relatively new. It’s target user are citizen scientists with moderate computer literacy. It has a low-to-medium learning curve.
The tutorial is still new. Please email me with feedback!
The oral annotation function is designed for the BOLD (Reiman 2010) methodology.

The annotation function for written transcription and translations is one of the easiest transcription tools to use. Written annotations are saved in eaf format (ELAN). You can doubleclick on the filename to open them up in ELAN and do more detailed annotations. Or export them to FLE, Toolbox, or other tools.
The progress chart shows the BOLD stages of the sessions. It also shows breadth of your corpus by genres.
How does SayMore clear up of metadata confusion? - Pre-labeled metadata fields are presented in attractive and easy-to-read fill-in-the-blank forms. The field labels can be translated into any language but everything is saved in standard XML format. Metadata can also be exported as a .csv spreadsheet.
How does SayMore clear up of metadata confusion? -
Metadata about participants can be linked to recording sessions. Informed consent files can be added to the people metadata. If a SayMore pre-defined file label is used, an indication will appear that Consent has been recorded. Otherwise, a yellow warning sign will remain – as shown in the next slide.
How does SayMore clear up metadata confusion?

SayMore alerts users to ethical issues by tracking where informed consent was (not) recorded. To help with access management – who can see what – SayMore gives access protocols from several archives. Once a protocol is chosen, access permission levels from that archive are presented as drop-down menu in the metadata “Access” field.
How does SayMore simplify file management?

SayMore does not store files in itself. It uses the folder directory of your PC. Each folder corresponds to an identically id’ed SayMore session. There is little chance of dragging and dropping files to the wrong folder.
How does SayMore clear up of metadata confusion?

- It manages file naming. SayMore does not allow two files or sessions to be given the same name. Once a session is id’ed, all files added to that Session are given the same id as a filename prefix. The user can customize the rest of the filename or use SayMore pre-defined filename labels. These pre-defined labels track the stages of sessions which is then reflected in the Progress Chart.
How does SayMore ease the archiving process? -

To archive, the user simply chooses a session or a project and clicks "Archive with...". Packages can be built for various metadata standards or customized to the preferences of specific archives. Currently only two packages are available. Feedback and cooperation with other archives could increase these packages.
I joined this group as the “documentation specialists. Files were located on various devices, with various amounts of metadata. They knew the importance of documentation and especially of archiving, but were too busy with MA theses and literacy projects to commit time to archiving. They had started to use SayMore and I was somewhat familiar with, so we decided to organize all data using SayMore.
Initially, recordists did not have time to answer metadata questions face-to-face. I was unsure how else to get the information. Instead, I asked them to fill out SayMore metadata forms whenever they had a few minutes to spare. They were able to record the metadata at their own pace and finished more quickly than anyone expected.

Initially, we planned to protect the file management from human error by creating a “partition” between material being archived and material that could be available for current work, but because SayMore makes it hard for users to drop files in the wrong place and does not allow sessions or files to be given the wrong name, we were able to open the whole collection of recordings on our server to anyone.

Initially, we wanted to archive at SIL since most of the linguists were SIL members and we wanted to archive somewhere more accessible to the community, who generally do not speak English. However, the most accessible archive required the metadata be in IMDI standard. The amount of work required to reformat the metadata to IMDI
nearly made us change our minds about archiving in the second place. However, after requesting an IMDI package from the SayMore developers, we are able to reformat the metadata with the click of a button.
A version with these new features should be released in 2015. For now, the “one-click” depositing is destined for SIL’s institutional repository only. As more archives dialog with the developers, the options will hopefully increase.

- Unified access protocol with automatic conversion to archive’s protocols during archiving process
- Pre-flight function
  + Selects sessions based on status
  + Checks for missing files
  + Give alerts about empty metadata fields
  + Generates a Table of Contents
- Deposits package at the click of a button
Perhaps you have more ideas?
What about building other software tools with similar features but filling their own niche?

**IDEAS FOR FURTHER IMPROVEMENT**

- Archiving packages for more archives
- Complete interoperability with Arbil and other programs
- More localizations/translations of the interface (currently: Russian, Spanish, French)
- Increase “hovering” explanations for metadata fields
- Obtain suggestions from archivists and from related disciplines (e.g. revitalization, literacy, anthropology, ethno-arts) about useful metadata fields
- Include stage markers that guide the documentation/description of language structures
- Add keywords/tags in metadata that indicate potential applications for sociolinguistics, language conservation, pedagogy, etc.
- …?