

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

METADATA

- ✘ Reminders of what to gather
- ✘ Variations and invented metadata labels should be limited
- ✘ Correctly using a metadata standard requires active knowledge of that standard
- ✘ Metadata should be in XML format
- ✘ Access management must be considered

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)

From Drude, Sebastian. 2011. File-Management: Organization, administration, and synchronization of files. Slideshow. Paper presented at the DOBES summer school Language Documentation, Regensburg.

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?

INTIMIDATING ARCHIVING PROCESS

- ✘ Finding and identifying the most appropriate archive for one's project
- ✘ Preparing corpus for depositing
- ✘ Satisfying the archive's preferences

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.

SAYMORE

The screenshot displays the SayMore software interface. The top menu includes Project, Session, Person, and Help. Below the menu are tabs for Project, Sessions, and People. The main window is divided into several sections:

- Sessions List:** A table with columns for Id, Title, Stages, and Status. The selected session is ENG-N01, titled "Life in Yorkshire".
- File List:** A table showing files associated with the session, including audio files and annotations.
- Session Details:** A form for editing session information, including ID, Date, Title, Setting, People, Location, Genre, Access, Situation, and Description.
- More Fields and Custom Fields:** Two tables for additional data entry.

Id	Title	Stages	Status
ENG-D01	Day and Night	■ ■ ■ ■ ■	○
ENG-E01	Unique Words	■ ■ ■ ■ ■	○
ENG-E02	Mirror Work	■ ■ ■ ■ ■	○
ENG-M02	Yorkshire Song	■ ■ ■ ■ ■	○
ENG-N01	Life in Yorkshire	■ ■ ■ ■ ■	●
New Session 01		■ ■ ■ ■ ■	○
New Session 02		■ ■ ■ ■ ■	○
SMTutorial_00	Word list	■ ■ ■ ■ ■	○
SMTutorial_04	Song	■ ■ ■ ■ ■	○

Name	Type	Date Modified	Size
ENG-N01.session	Session	2/14/2015 11:35 ...	1 KB
ENG-N01_Source.wav	Audio	2/14/2014 1:13 AM	508 KB
ENG-N01_Source.wav.annotations.eaf	Annotations	5/31/2014 2:12 PM	2 KB
ENG-N01_Source.wav.oralAnnotations.wav	OralAnnotations	5/31/2014 12:45 ...	2.81 MB
ENG-N01_Source.wav.oralAnnotations.aup	Audacity Proj...	12/17/2014 4:21 ...	2 KB

Field	Value
Researcher Involvement	Elicited
Location Country	United Kingdom
Location Continent	Europe

Field	Value
Researcher	
PreviousName	
PreviousLocation	

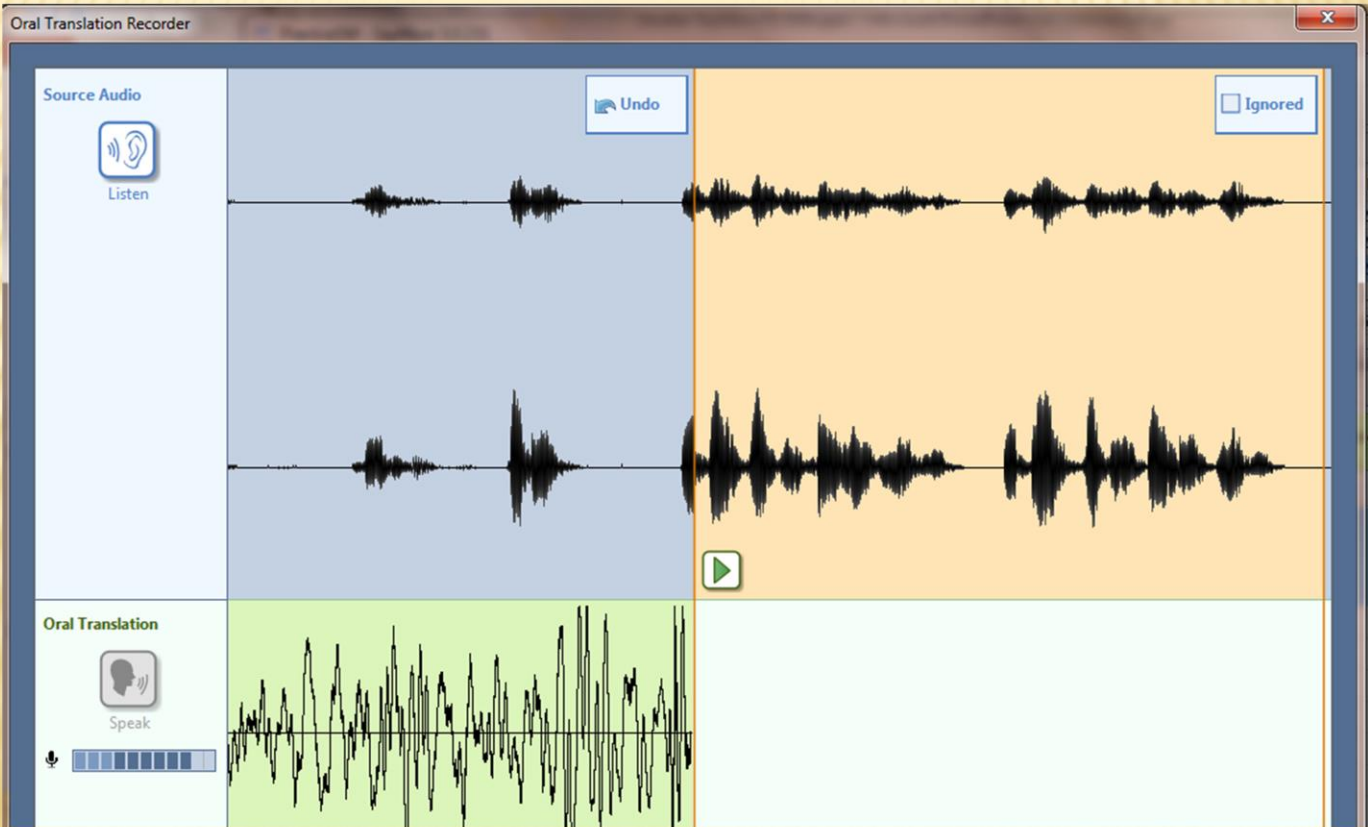
SayMore is an SIL software, still relatively new. It's target user are citizen scientists with moderate computer literacy. It has a low-to-medium learning curve.

SAYMORE

- ✘ Download: <http://saymore.palaso.org/>
- ✘ Tutorial: <https://drive.google.com/open?id=0B2TFy02zqdRwOE9sdEZaNFipcHM&authuser=0>

The tutorial is still new. Please email me with feedback!

ORAL ANNOTATIONS



The oral annotation function is designed for the BOLD (Reiman 2010) methodology.

Reiman, D. Will. 2010. Basic oral language documentation. Language Documentation & Conservation. 4.254-268

*<https://scholarspace.manoa.hawaii.edu/bitstream/handle/10125/4479/reiman.pdf>
[accessed: March 4, 2015]*

WRITTEN ANNOTATIONS

ENG-M02_Source.WAV	Audio	5/31/2014 2:03 PM	24.21 ...	00:01:35
ENG-M02_Source.WAV.annotations.eaf	Annotations	12/17/2014 3:20 ...	9 KB	
ENG-M02_Source-ELAN.flextext	FLEXTEXT File	6/12/2014 10:36 ...	25 KB	

Annotations

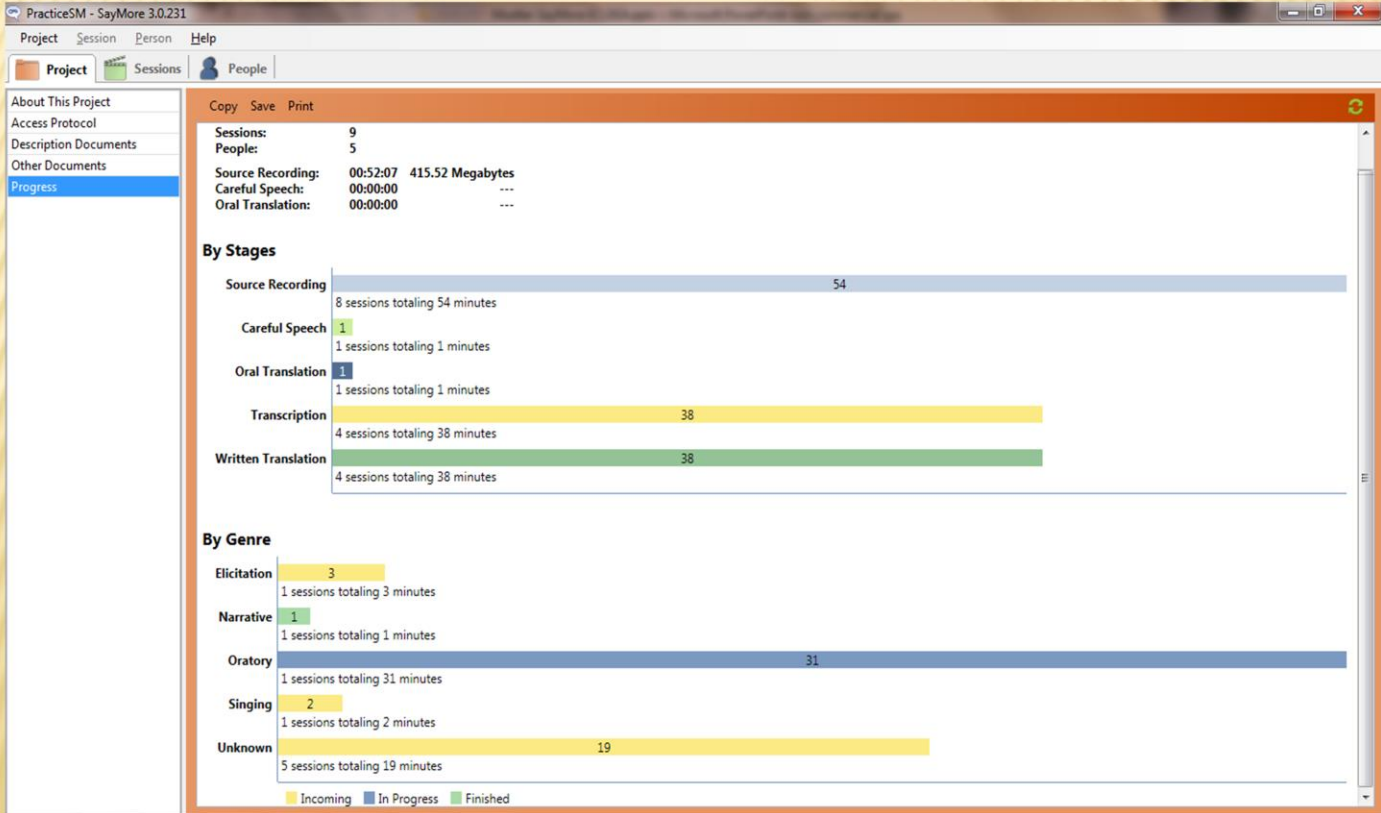
100% Oral Annotations Tools Segment... Export ?

	Transcription	Options	Free Translation
1	Ignored		
2	Wheear 'ast tha bin sin' ah saw thee?		Where have you been since I sa
3	On Ilkla Moorar baht 'at		On Ilkley Moor without a hat
4	Wheear 'ast tha bin sin' ah saw thee?		Where have you been since I sa
5	Wheear 'ast tha bin sin' ah saw thee?		Where have you been since I saw you:
6	On Ilkla Moorar baht 'at		On Ilkley Moor without a hat
7	On Ilkla Moorar baht 'at		On Ilkley Moor without a hat
8	On Ilkla Moorar baht 'at		On Ilkley Moor without a hat
9	Tha's been a cooartin' Mary Jane		You've been courting Mary Jane
10	On Ilkla Moorar baht 'at		On Ilkley Moor without a hat
11	Tha's been a cooartin' Mary Jane		You've been courting Mary Jane
12	Tha's been a cooartin' Mary Jane		You've been courting Mary Jane
13	On Ilkla Moorar baht 'at		On Ilkley Moor without a hat
14	On Ilkla Moorar baht 'at		On Ilkley Moor without a hat
15	On Ilkla Moorar baht 'at		On Ilkley Moor without a hat

- Audacity Label File (Free Translation)...
- Audacity Label File (Transcription)...
- ELAN File...
- FLEX Interlinear Text...**
- Plain Text...
- Subtitles File (Free Translation)...
- Subtitles File (Transcription)...
- Spreadsheet (CSV) File...
- Toolbox File...

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 FLEx, Toolbox, or other tools.

PROGRESS CHART



The progress chart shows the BOLD stages of the sessions. It also shows breadth of your corpus by genres.

SAYMORE METADATA

Session Status & Stages Notes

Title	Mirror Work	Setting	Echoey room, small room with nothing on walls, Sarah interrupting, Will "trying" to help
People	Shakespeare, William	Location	Key 300 ILC
Genre	<Unknown>	Access	U
Situation	small work group in workshop standing around, setting up, learning, we are dealign with new equipment	Description	Swadesh word 22-33. Tones hummed after each word. English translation given immediately before.


More Fields		Custom Fields	
Field	Value	Field	Value
Researcher Involvement		Researcher	
Location Country		PreviousName	
Location Continent	Elicited	PreviousLocation	
Location Region	Non-elicited		
Location Address	No-observer		
Planning Type			
Sub-Genre			
Social Context			

Investigator asks speaker(s) to produce isolated phonemes/ words/ utterances / grammatical structures.

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

SAYMORE METADATA

Person Contributions Notes

Full Name Jonathan Yorke	Birth Year 1988	
Nickname Jonny	Code	
Primary Language Yorkshire SL	Gender Male	How to Contact 8-888-888-8888 jonathan_yorke@mji.org skype: j.yorke
Learned In: Leeds		
Other Languages French LRW		
Russian SLRW		
Queens English SLRW		Ethnic Group English
English Variety B		Primary Occupation Boompole operator
Education MA Linguistics Middlesex University (English)		

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.

SAYMORE METADATA: ACCESS MANAGEMENT

Person	Consent
Jonathan Yorke	
P01	
New Person 01	
F02	
Shakespeare, Willi...	

No Informed Consent

Access Protocol used by this project: **ELAR** [Help for access protocols](#)

- None
- AILCA
- AILLA
- ANLA
- ELAR**
- REAP
- TLA
- Custom

Choices:

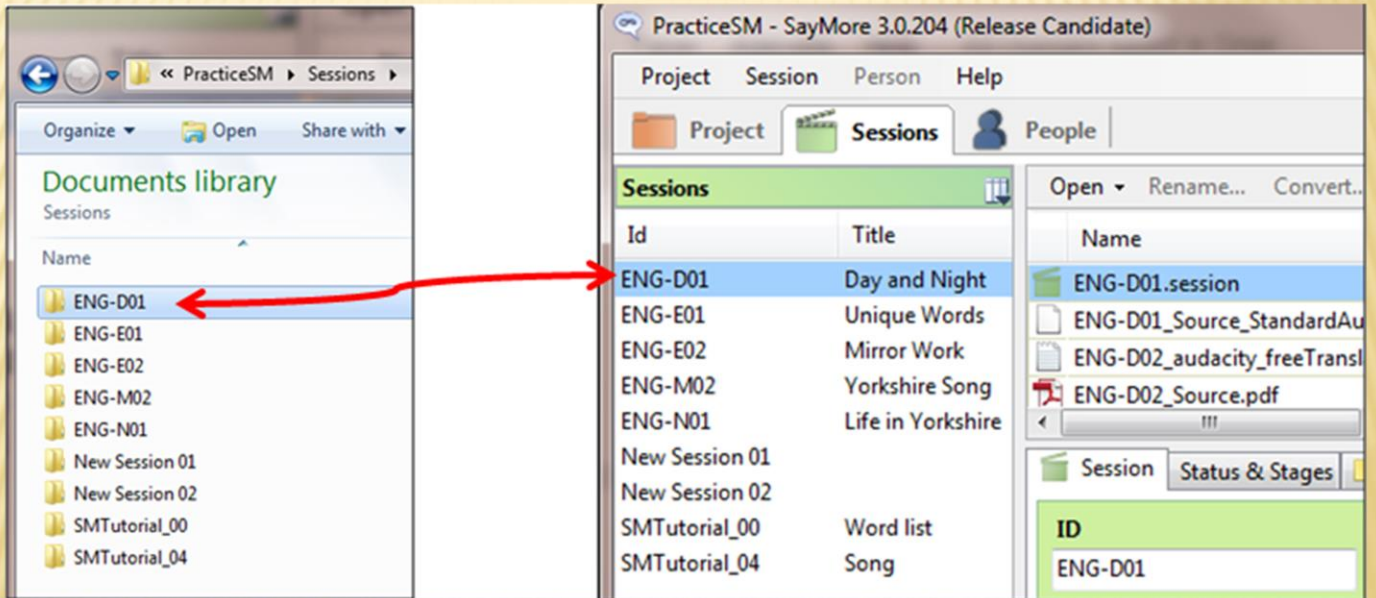
- U - all Users can access
- R - Researchers and Community members are allowed access
- C - only Community members are allowed access (normally requires application to Depositor)
- S - only Subscribers are allowed access (requires application to Depositor)

How does SayMore clear up of 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.

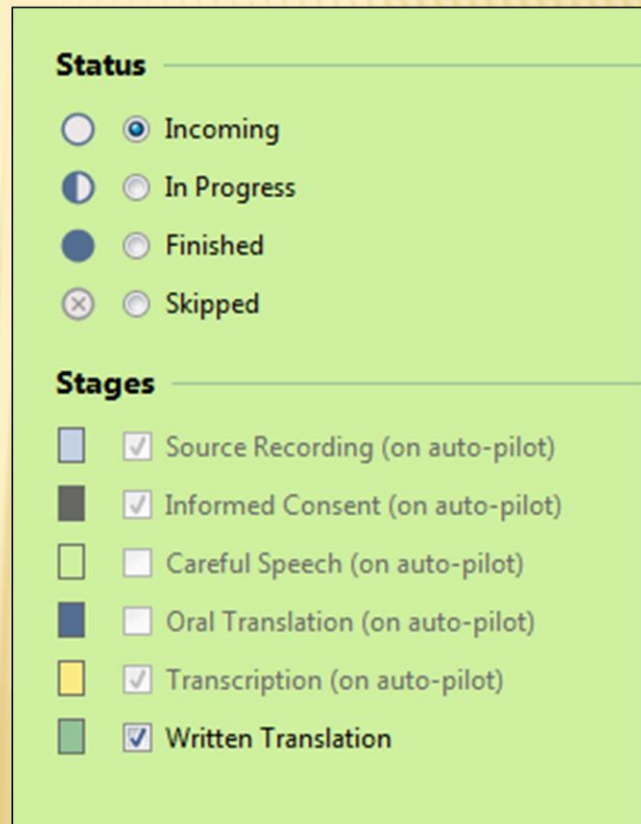
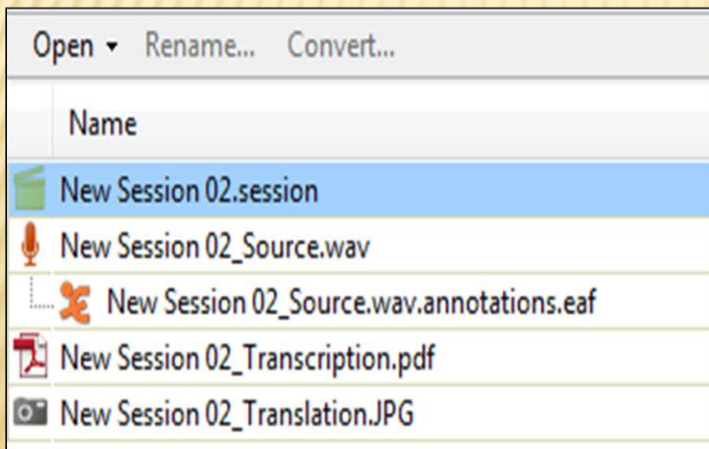
SAYMORE FILE MANAGEMENT



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.

SAYMORE FILE MANAGEMENT



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.

SAYMORE ARCHIVING

- ✘ Creates archiving packages
 - + SIL's RAMP (for it's institutional repository)
 - + IMDI (can be imported to Arbil or LAMUS)
- ✘ Packages whole project or single session

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.

CASE STUDY

- ✘ Rutul and Tsakhur, Nakh-Daghestanian languages spoken in Russian and Azerbaijan
- ✘ Linguists and some community members
- ✘ Only one person familiar with language documentation
- ✘ Since 2009, over 500 audio and video recordings (~11 hours) had been gathered
- ✘ They “don’t know where to start” on archiving

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.

CASE STUDY

- ✘ Fill-in forms and attractive interface engaged recordists to enter metadata and maintain consistency
- ✘ File management “safeties” opens the collection to anyone
- ✘ Multiple archiving packages encouraged archiving with accessibility in mind

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.

IMPROVEMENTS UNDER CONSIDERATION

- ✘ 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

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.

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.
- ✗ ...?

Perhaps you have more ideas?

What about building other software tools with similar features but filling their own niche?