

Slide 1



The Archive of the Indigenous Languages of Latin America

Simple Steps for Archiving Language Documentation Data

Susan Smythe Kung J. Ryan Sullivant Elena M. Pojman

Archive of the Indigenous Languages of Latin America
University of Texas at Austin
January 3, 2020
SSILA

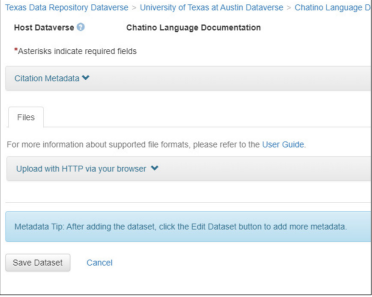


Slide 2

Participatory Archiving in Language Data Archives

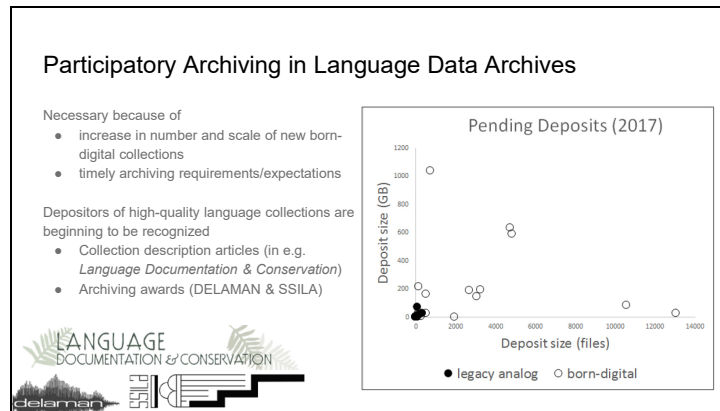
Depositors now do more archival tasks

- Appraisal
- Arrangement
- Metadata creation + entry
- Collection description
- Media ingestion



The screenshot shows a web interface for 'Chalino Language Documentation' within the 'Host Dataverse' system. The page title is 'Chalino Language Documentation' and it includes a breadcrumb trail: 'Texas Data Repository Dataverse > University of Texas at Austin Dataverse > Chalino Language Documentation'. A note states '*Asterisks indicate required fields'. The form is titled 'Citation Metadata' and includes a 'Files' section with a text input field. Below this, there is a link to the 'User Guide' for supported file formats and an 'Upload with HTTP via your browser' button. A 'Metadata Tip' box suggests clicking the 'Edit Dataset' button to add more metadata. At the bottom, there are 'Save Dataset' and 'Cancel' buttons.

Language data archiving has become participatory, meaning that depositors themselves are performing tasks that were previously done by language archive staff. Depositors can now expect to appraise their materials to determine what gets added to a collection, how those materials will be arranged, create and enter different kinds of metadata -- from basic citation metadata to rich descriptive metadata, technical metadata, and rights management metadata. Depositors are also more often than not responsible for writing holistic descriptions of their projects and collections, and in some repositories, they will ingest their media files themselves through graphic interfaces like the one seen here.



Now, this participatory turn was sorely needed in the domain of language data archiving since there are more collections now that are being preserved (thanks to archiving requirements and changing researcher attitudes towards archiving) and today's born-digital collections are much bigger than the digitized analog collections of yesteryear. On the right is a chart I made to learn about the backlog I inherited. All the legacy analog collections (the filled circles) had fewer than 500 files and took up less than 100 GB, while most collections with born-digital material (the empty circles) were bigger--often much bigger--in one or both of these dimensions.

The depositor's role and contribution to this participatory archiving is also starting to be recognized within the field. The journal LD&C is publishing collection description articles at a steady clip now, and there are now two bodies giving awards to depositors for archiving notable collections of materials.


Participatory Archiving in Language Data Archives

Advantages:

- Familiarity with their materials → better descriptions
- Familiarity with languages in their collections
- Timely archiving & updating

Challenges:

- Archiving skills are outside of training
- Depositor-curated collections often fail to include critical metadata (Koshoffer et al., 2019; Grant et al., 2019)



Collectors often create incomplete descriptions.

Depositors actually have some advantages over archive staff when it comes to arranging and describing their collections. First off, depositors have a great deal more knowledge about the collection's context that would take archive staff a considerable amount of time to reproduce, if they even can. Secondly, depositors taking over many of the collection processing steps reduces bottlenecks and backlogs in the repository leading to more timely archiving and updating of collection materials and catalog records.

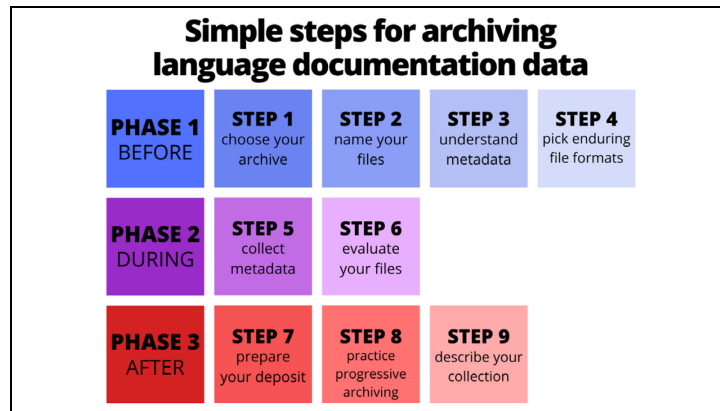
However, there are some challenges. Archiving tasks do take some time and attention, and are not part of many researchers' training. This is a problem faced by many kinds of research data archives -- depositor curated collections often fail to include the kinds of rich contextual descriptions that depositors could create. Some studies found that more than half of depositor-curated datasets did not clearly identify what the data was nor how it was generated. This is of course understandable, since unless we are trained otherwise, we write descriptions and metadata for ourselves, not the unknown people who will run across these materials later.

Depositor Instruction

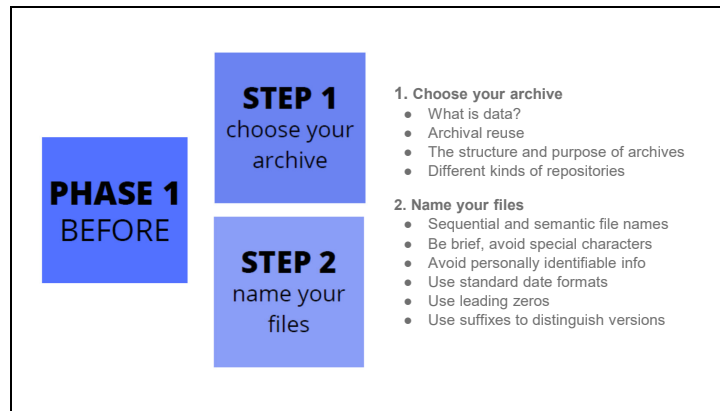
Supported by an NSF-DEL grant (BCS-1653380) to create a repository-neutral curriculum to guide and advise language data depositors

- Interviews with language and data archive managers
- Environmental scan + literature review
- Courses at Colang 2018 (*with Vera Ferreira and Alicia Niwagaba*) + Colang 2020 (*with Jaime Pérez González*)
- Short video tutorials (*with Alicia Niwagaba, Nishmet Montelongo, and Judith Lara*)
- Journal article under review at *Language Documentation & Conservation*
- Online learning modules (*with Elena M. Pojman*)

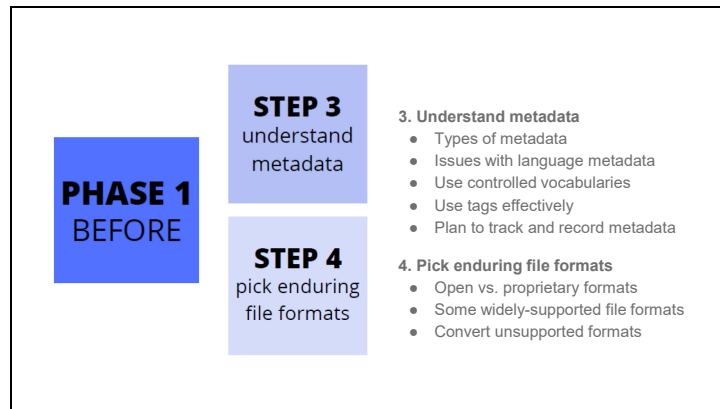
To address this skills gap, we undertook an NSF-DEL funded project to build a repository-neutral curriculum to guide and advise language data depositors. Our research involved interviewing managers of language and data archives, performing an environmental scan, and doing a literature review. The curriculum was offered at Colang 2018 with our colleague Vera Ferreira and will be offered again at Colang 2020 with our colleague Jaime Pérez González. We also produced short video tutorials of some topics with our graduate research assistant Alicia Niwagaba and undergraduate intern Judith Lara, and today we'll be talking about the online learning modules that Elena Pojman, our undergraduate research assistant is helping us build.



We've organized our curriculum into nine steps that can be grouped together into three phases: planning steps to be taken before data collection begins, steps to be taken while collecting data, and steps to be taken after data collection and before submitting materials to an archive. While we present these materials in a linear order, some of these activities are iterative or may have to be revisited later.

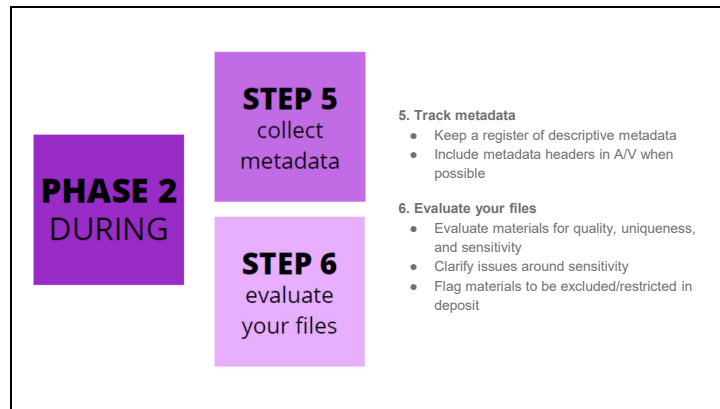


Before data collection, researchers should ideally choose which archive they will be depositing their data with, as this will help inform many other choices made throughout the process. In this step we also help researchers understand the purpose of data archives, provide some examples of how archival collections are used, and discuss different kinds of repositories. In the next step, we teach researchers best practices for creating file names and offer some examples of file naming schemes that have worked for researchers and archives alike.



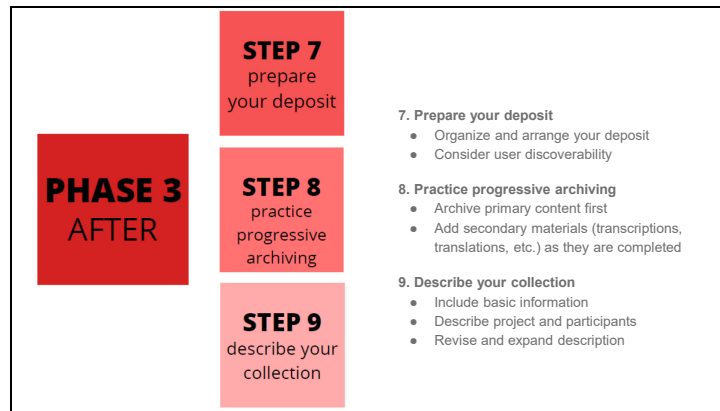
In the next step, we explain issues around the different kinds of metadata they will be responsible for recording. In particular, we offer advice about tracking and reporting language metadata, interpreting repositories' controlled vocabularies, and advice on effectively using open-ended keyword tagging systems.

In step 4, we explain how to select appropriate archival file formats for the collection and provide some information about some widely supported archive-ready formats for text, images, and A/V media and how to convert non-archival files to these formats.



The second phase covers actions that should be taken iteratively throughout the data collection process. Step 5 covers collecting the metadata categories identified back in step 3 for every set of digital objects and--whenever possible--including metadata at the beginning of AV files themselves.

Step 6 teaches researchers to continually evaluate their materials for their quality, uniqueness, and sensitivity to flag those materials that should be kept out of the deposit or given special treatment. We urge researchers to do this work while still in their fieldwork context so that remedies can be made (e.g. rerecording an event or a respeaking of a noisy audio file) or issues around access to sensitive materials can be clarified

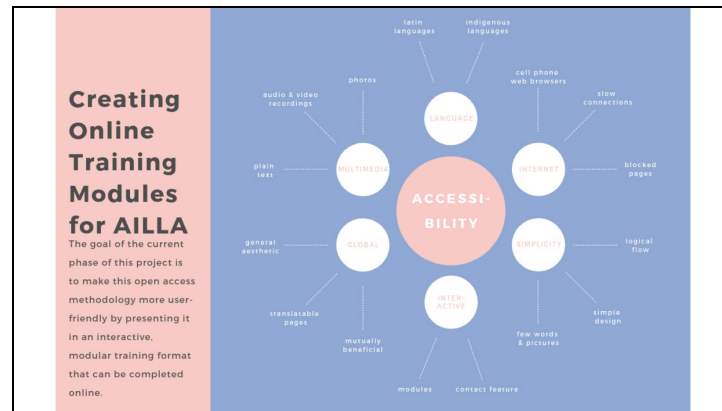


The final phase covers what can be done to prepare a collection for deposit at an archive. Step 7 teaches researchers to organize and arrange their deposits to accommodate data archive structures and enhance discoverability. Step 8 explains the iterative process of progressive archiving for growing data collections, and step 9 provides advice and guidance for creating collection descriptions that provide a full account of the context of a collection's creation, its structure, and how to find and use items within it.

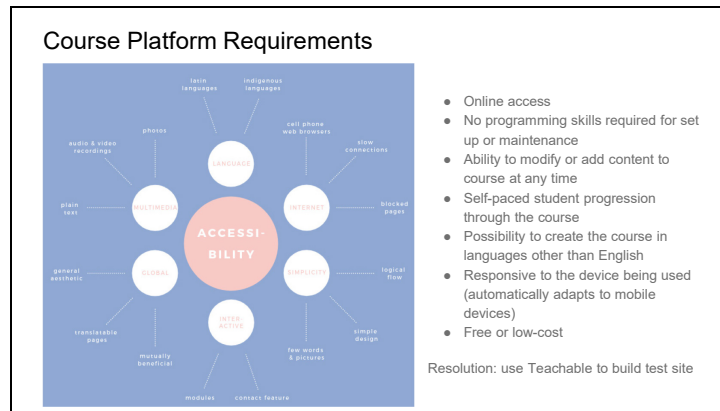
Slide 11



After Kung et al. taught the first version of this workshop at CoLang 2018, Kung had the idea to create an online course that would be accessible for anyone to complete as they had the time and the need to do so. This year we are building a proof of concept that we plan to release later this year.

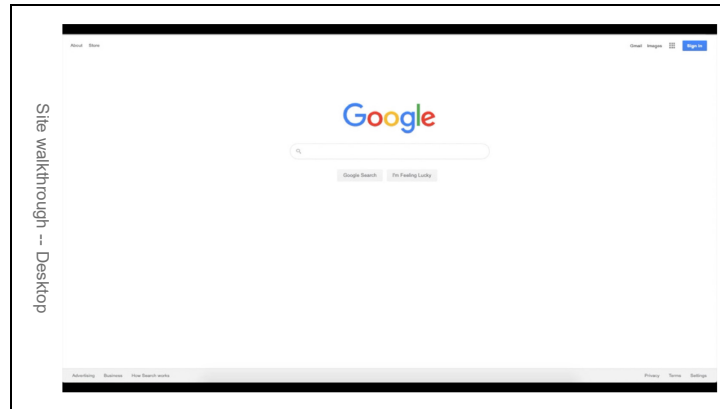


When I began the project, I decided to identify key features that our training program would need to include. I made it my goal to make it as accessible as possible, so that anyone, anywhere could access our course and benefit from it. Thus, I listed key concerns that would be vital to increasing accessibility to our course. It needed to be an internet-based program so people could log in from anywhere at any time, and it needed to be responsive, meaning they could access it from any device. It couldn't be blocked in countries where we may have people collecting data. It must be beneficial to people with slow connections and limited space on their devices. It must work for characters beyond just English and even Spanish. It must allow for us to use a variety of mediums to transmit information -- plain text, images, audio, and video. The site must be simple to use and view, and must not require advanced knowledge of computers. It must be interactive, allowing for users to move through lessons that flow logically. Lastly, it must fit a global aesthetic that lets users know that our relationship is mutually beneficial.



We built an example course on the platform Teachable, which we will present today. We are still in the process of finalizing the platform that we will use for our final version. However, Teachable included most of the functionalities that we needed, and worked well as a test site.

Slide 14



(0:00 - 0:13) google home screen leading into Archiving for the Future home screen

Let's navigate to the test version of our site, titled "Archiving for the Future". There will be an English and Spanish version, and potentially a Portuguese version as well. Let's look at the English version now.

(0:14 - 0:28) overview of curriculum view on English page

The English-language home page shows an overview of the entire curriculum. The "simple steps" shown before in the infographic are reflected in the phases and lessons.

(0:29 - 0:48) explanation of curriculum icons

At the left-hand side, each lesson type is represented by an icon that describes its contents. Text pages show a piece of paper, quizzes show a light bulb, and videos show a play button.

(0:49 - 0:57) login details

Now let's preview the curriculum. Creating an account requires your name, email, and a password, which preserves your progress through the course.

(0:58 - 1:22) class page

Once enrolled, you will be taken to your courses. You can see the courses in which you are enrolled, and your progress through them. Within the course you can see where you left off and a glance at the remaining lessons, broken down by lesson type and lecture. This page shows your instructor as well. Let's take a look at the curriculum.

(1:23 - 2:11) explanation of icons + progress bar

This first lesson gives an overview of our “simple steps”, represented in our lessons. At the top left-hand side, you can see your progress through the course, which increases as you press the “complete and continue” button at the top right-hand side. As lessons are completed, the corresponding lesson on the left change from an open circle to one filled, with a check mark inside. Lessons that have been started but not yet completed are half filled. Lessons vary in length and medium.

On the left, you can track your progress through the course and see the remaining lessons and sections.

(2:12 - 2:42) quiz run-through

Within the course we have created quizzes as a refresher of the most important concepts. You can immediately see if you were correct or incorrect. Quizzes with multiple correct answers will alert you if you did not select all correct solutions.

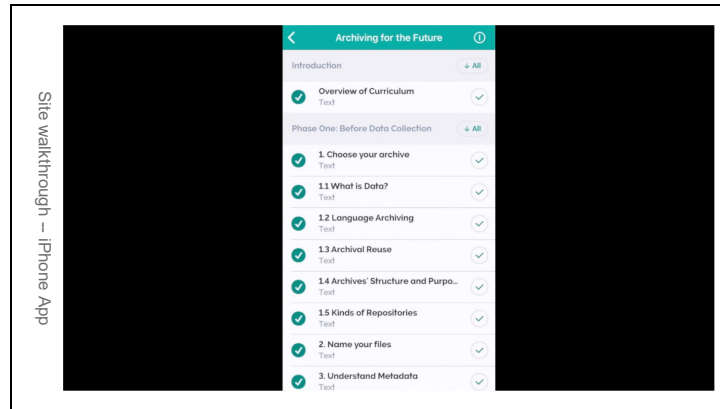
(2:43 - 3:05) sample lesson

Here is a sample lesson, which is mixed media: it includes text, images, and video.

(3:06 - 3:14) last look

Here we see the course curriculum again, which we saw when we logged in. It updates as we progress through the course.

Slide 15



(0:00 - 0:12) overview of curriculum on iPhone app

Now, let's take a look at the course in the Teachable iPhone app. The course is optimized for your device, and can be downloaded to be accessed from anywhere offline.

(0:13 - 0:22) look through lessons + completing lessons

You can scroll through the curriculum easily, and complete and continue lessons by pressing the button at the bottom.

(0:23 - 0:31) finishing course

When you have completed every module, Teachable recognizes your effort!

Thank you!

Contact:
ailla@ailla.utexas.org
@AILLA_Archive

Videos Playlist: <http://bit.ly/ArchiveLgDoc>

References

- AILLA. 2018. Linguistic Data Curation Tutorials. Archive of the Indigenous Languages of Latin America. Access: Public.
<https://ailla.utexas.org/information/object/ailla/257379>
- Digital Endangered Languages and Music Archives Network. 2019. DELAMAN Award. <http://www.delaman.org/delaman-award/>
- Grant, Rebecca, Graham Smith, & Iain Hrynaszkiewicz. 2019. Assessing metadata and curation quality: A case study from the development of a third-party curation service at Springer Nature. *bioRxiv* 530691. doi: <http://dx.doi.org/10.1101/530691>
- Koshofer, Amy, Amy E. Neeser, Linda Newman, & Lisa R. Johnston. 2019. Giving datasets context: A comparison study of institutional repositories that apply varying degrees of curation. *International Journal of Data Curation* 13(1), 15-34.
<http://dx.doi.org/10.2218/ijdc.v13i1.632>
- Society for the Study of the Indigenous Languages of the Americas. 2019. Archiving Award. <https://www.silla.org/awards/archiving/>

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. BCS-1652880. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.