Constructing Online Language Learning Content Archives for Under-Resourced Language Communities[*]

Russell Hugo

Department of Linguistics, University of Washington,
Guggenheim Hall 4th Floor, Box 352425, Seattle, WA 98195-2425
rlhugo@uw.edu

Summary: This paper describes a case study for the creation of an archive of learning materials for the Sahaptin language. Between approximately 1990-2003, Dr. Virginia Beavert, a Yakama elder and native speaker of Sahaptin, and one of her students, the late Edward James, taught Sahaptin language classes at what is now Heritage University in Toppenish, WA. The materials developed for that class were the focus of this project. In addition to examining the processes involved with the sorting, digitization and organization of paper materials, the paper looks at issues relating to multimedia, collaboration, curation, and security. The final product involved an online archive that was intended to maximize community access and be duplicatable for similar communities with limited resources. At the same time, the archive attempts to address best practices for long-term archiving, including XML metadata.

Keywords: endangered languages, archiving, materials, Sahaptin

[*] Thank you to Virginia Beavert, Mary James, the UWWP in Linguistics and the anonymous reviewers for their helpful comments, Prof. Emily Bender, Prof. Sharon Hargus, as well as Theodore Gerontakos and Dan Mandeville for their generous oversight and suggestions regarding the metadata template.

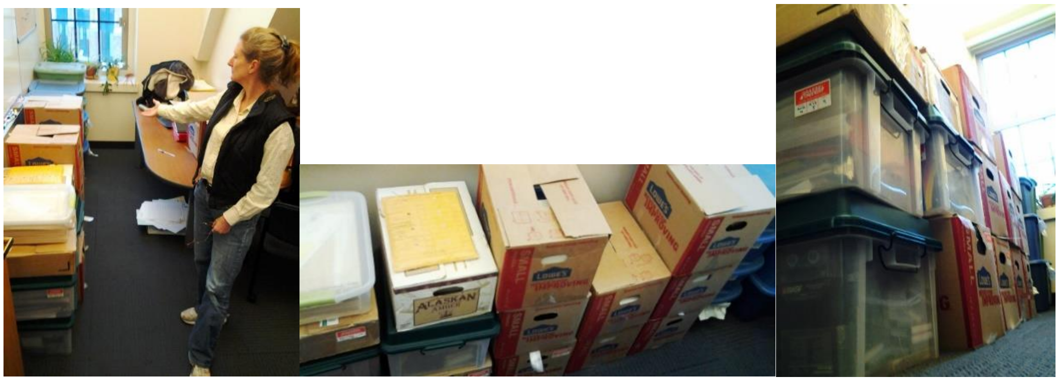
University of Washington Working Papers in Linguistics (UWWPL). Volume 33 (2015).

1 Background

Origin of the materials

Between approximately 1990-2003, Dr. Virginia Beavert, a Yakama elder and native speaker of Sahaptin, and one of her students, the late Edward James, taught Sahaptin language classes at what is now Heritage University in Toppenish, WA. For those classes Dr. Beavert, with the help of Edward James, who eventually became an instructor of Sahaptin, developed numerous pedagogical materials, which were collected and stored by Edward James. Edward James died in 2011, and in 2013, Mary James, his widow, gave the materials to University of Washington Professor Sharon Hargus for preservation and archiving, who provided me the opportunity to work with them. The materials consisted of roughly 21 banker boxes containing paper materials (**Figure 1**), over 300 audio cassette tapes and a handful of video cassettes. Under the supervision of Prof. Hargus, I began preparing the paper materials for archiving. After the materials had been digitized, due to the UW archives not having the means to construct a community accessible archive at that time, it was necessary to develop one that could provide access in the short term but allow a smooth transfer to a permanent archive. An online archive system was deployed which was constructed around the common and well-supported Content Management System (CMS) Wordpress. In this paper, I will outline some of the processes used and challenges faced when constructing a short-term and accessible archive that is feasible for a less-resourced community to build and manage. The specific archive this chapter is concerned with is intended to be housed in a permanent archive in the near future and must be readily transferable when one is found.

Figure 1: Materials as delivered to the UW (Left: Prof. Sharon Hargus discussing the newly arrived materials.)



Challenges

From the outset, there were two challenges facing this project:

1) How can the materials and the information they contain be preserved for as long as possible?
2) How can the materials be made as accessible as possible (respecting necessary restrictions)?

The question of preservation is, in and of itself, a common focus of discussion, and Borghoff, Rodig, Scheffczyk, & Schmitz (2005), in particular, provide a good review of the issues there in. Physical media has traditionally had a very limited lifespan, and using computer-based physical media (such as floppy disks and CD-ROMs) to archive material has become problematic in a remarkably short time. In 1995, it was considered perfectly reasonable to archive documents in the

Word Perfect (Proprietary Software) format on 4.25” floppy disks, but in less than 20 years, the changes in computing have made accessing media in that format difficult (assuming the data is still viable and the physical media, itself, has not been corrupted). Newer forms of digital storage seem to be more promising (e.g., cloud-based) but present new challenges as well.

From the outset a distinction needs to be made between the temporary archive, which will be the focus of this chapter, and an eventual permanent archive. The Consultative Committee for Space Data Systems (CCSDS) Open Archival Information System (OAIS) (CCSDS, 2012) uses ‘long term’ to describe any archive that is intended to be kept, hypothetically, forever. The archive system that was constructed for these materials does not meet the standards outlined by CCSDS and the infrastructure to support it for the long term was not currently available. Therefore, the constructed archive is a ‘temporary’ archive intended to improve the accessibility of the materials in the short term. Eventually, the materials will be transferred to a ‘permanent’ home, which is an archive that is equipped for ‘long term’ storage. ‘Short term’ can be defined as a period as long as a community can support it. As will be discussed later on, the infrastructure and resources required for ‘long term’ archiving are substantial and beyond the means of most, if not all communities. Therefore, it is impossible to clarify or predict the length of the term for a ‘short term’ archive. Instead, the utility of the term is to emphasize the caution and maintenance that will be necessary should a community forgo having content held in a ‘long term’ archive. This issue will be explored further throughout this chapter.

As such, the permanent archive where the Sahaptin materials will eventually be stored should be equipped to deal with technological changes, making it imperative for the temporary archive to store materials in a way that reduces the risk of loss or corruption of data as much as possible. “Preservation is not a discrete process, but rather a never-ending management task” (Edmondson, UNESCO & Information Society Division, 2004:20). The paper materials, which I focus on in this chapter, have an arguably longer lifespan than most digital media when said paper materials are stored in a controlled environment, but they, too, are at risk for loss or damage (e.g., fire, water, mold, etc.). The issue of long term preservation and accessibility, therefore, becomes of paramount concern and will be discussed further in (§3).

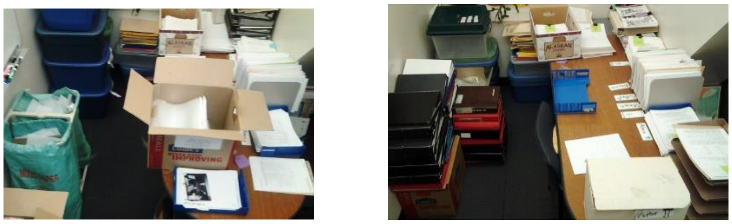
Sorting

The first step in archiving the Sahaptin material involved assessing the materials present. I worked with Prof. Hargus to ensure a large portion of the boxes were opened and that an initial audit of types and quantities was done. Much of this work required Prof. Hargus’ language expertise, and from this audit, we developed an initial set of categories into which we sorted the materials. A makeshift sorting station was created in Prof. Hargus’ office with receptacles representing each of the initial categories (e.g., culture, grammar, lexicon) and a larger bin was added for materials that didn’t fit into one of the existing categories. Phase 1 of the sorting began with the removal of unnecessary coverings (e.g., plastic sheets and binders) and included a careful recycling of clear duplicates. Many of the documents contained only minor differences, which made this step of the process an arduous task. An assignment given five years in a row, for example, might contain slight amendments each year or a different set of hand written annotations, so when any of the content differed, all versions were kept for archiving. In instances where content was wholly identical, only the copy with the highest quality and most legible text was retained.[1]

[1] Exact duplicate copies were not seen as valuable for this project as their existence in the original material set were not duplicated for archival purposes, but only due to limited organization. Since this was not intentional, information about how many copies of a certain version of a certain materials provided no value and would add unnecessary clutter to the archive listing.

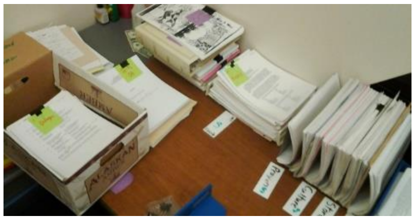
During this initial phase, Prof. Hargus reviewed the uncategorized materials and suggested new categories to be added as needed. Some of the materials didn’t belong in the learning material archive at all (e.g., random notes about the language), so these were set aside to be reviewed by Prof. Hargus at a later time.

Figure 2: Photos of the Phase 1 Sorting Process (L= Duplication check and recycling. R=Phase 1 completed.)



After all of the boxes had been reviewed once, the second phase of sorting began. Because many of the materials fell into multiple categories (e.g., grammar AND culture) each category of materials was reviewed for duplicates and re-categorized as necessary.

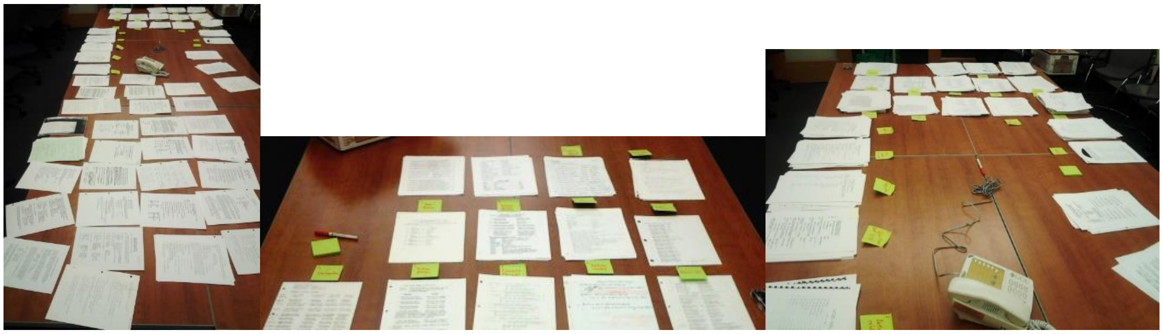
Figure 3: Photo of the Phase 2 Sorting Process



Each of the larger categories were then sorted a third time. This phase of sorting had three goals: 1) to eliminate further redundancies (i.e., duplicates); 2) to sort the material into relevant sub-categories; and 3) to organize relevant content by date or proper sequence, as some of the associated materials were spread across multiple boxes.

Figure 4, below, shows the sorting spread of a single category, so it is little wonder the sorting process took approximately four months of part-time work. We knew the better organized the materials were prior to scanning, the less room for error there would be when the scanning, bundling and archiving of the digitized versions occurred.

Figure 4: Photos of the Phase 3 Sorting Process



Digitizing the Text Materials

Selecting a format

Johnson (2004) discusses the three primary formats used in an archive—*archival, working* and *presentation*—and provides the following example: if you have a grammar, the *archival* format might be Extensible Markup Language (XML), the *working* format PDF/HTML, and the *presentation* format Microsoft Word. Since the source of all text materials in the Sahaptin Archive were physical papers, though, this distinction did not apply as strictly. It would have been possible to use an uncompressed archival format (e.g., TIFF), but size considerations and lack of proper equipment prevented this from happening. In conjunction, the materials were to be archived in the same form in which they were given to the UW, meaning many of the documents contained errors or had aesthetic issues. Simply reconstructing each document in an editable Unicode or vector format was not feasible. This created archival and working formats with identical but potentially inaccurate content. Thus, only the presentation format could be supplemented with notes on each material's dedicated page in the short term archive system.

According to Johnson (2007:7), the "general requirements for archival-quality (master copy) formats are that they be:

- non-proprietary; that is, their encoding is in the public domain;
- thereby amenable to forward migration to new formats over time;
- portable, re-useable, repurposeable;
- the best possible reproduction of the original (if not original themselves.)"

Borghoff et al. (2005) follow a criteria similar to that of Johnson, above, recommending standard formats for archived images (e.g., PDF, PostScript, TIFF, GIF and JPEG). While the TIFF format offers much better data preservation than PDFs, the extreme storage costs, the inability to easily bundle them without complicated processing (i.e., for multiple pages) and the display limitations (e.g., cannot be easily rendered in a browser) make TIFFs less practical for the purposes of this particular archive. Similarly, the GIF and JPEG formats, while browser friendly, lack Optical Character Recognition (OCR) compatibility as well as the bundling and print functionality of PDFs. PostScript offers no real advantages to PDF, either, other than in its compression differences.

The PDF standard is one of the most commonly used formats for documents today, and although technology is unpredictable, it is likely to have substantial longevity and, more importantly, portability in the future. A major downside to this digitization process is that the end result is an image and not a digital document. While PDF has some OCR capabilities for English, it is still problematic, and even more so when used with other languages. It was essentially useless with the Sahaptin character set. As a result, PDFs may require the sacrifice of layout and design elements, but an accompanying archival format in XML and/or RTF greatly improves the portability and usability of the data (Cushion, 2004). Doing such with the Sahaptin Archive, though, would have been far beyond the current state of OCR and would have been unfeasible with the limited resources available for the project. A compromise, therefore, needed to be struck. The present utility of the limited English OCR, while beneficial if an effort were to be made to fully digitize the text in the future, is primarily limited to being indexable by a search engine for improved accessibility and potentially speeding up the migration of the content to a text-editable format (e.g., DOC) by copying and pasting.

Adobe still has ownership of some of the functionality of the PDF format, but the functionality that was required by the materials in the Sahaptin Archive falls under the open source International Organization for Standardization (ISO) PDF format. The format that was selected needed to meet the non-proprietary requirements. PDF did not have the same type of oversight and level of standardization as the American National Standards Institute (ANSI) or the World Wide

Web Consortium (W3C), but "PDF…has become a de facto standard for the exchange of printable documents" (Borghoff et al., 2005: 36) and because of the scans' limited OCR support, they must be considered primarily images and not digital documents, or as Borghoff et al. term it, *hybrid* PDF documents.

Scanning

The PDFs of the Sahaptin materials were scanned on a XEROX ColorQube 9201[2] set to *high quality* and 300 dots per inch (DPI) resolution. The minimum amount of compression the copier allowed was used. This resolution was considered sufficient for the archive as the majority of the source documents were third (or more) generation photocopies. Testing found no OCR advantages for increasing the resolution beyond 300dpi and little visual improvement considering the accompanying file size increase. Additionally, print comparisons of the scanned PDFs and the original documents showed no substantial differences in quality. Increasing the resolution, even to 450dpi, added a substantial load on the network from the printer and increased the final size of the archive considerably. Therefore, the 300dpi standard was chosen to maximize the portability and longevity of the archive as well as to maintain a reasonable estimated time of completion.

Deskewing

A balance was sought between the quality of the scans and the efficiency of the process. Some bulk scans, for instance, resulted in a moderate skewing of the document, and other documents in their original form which were askew were left uncorrected. Adobe Acrobat does provides some deskewing functionality, but deskewing, in this case, would have resulted in the document being compressed again and losing further information. According to available documentation, the only way to bypass the compression would have been to extract each page into an uncompressed format (e.g., TIFF), import it into a photo editing suite (e.g., Adobe Photoshop) manually deskew it, save the file again, and then reconstruct the PDF. Doing so would have greatly increased the workload and the file size of the PDF if further compression was not then completed. For this reason, we decided that unless there was a loss of information, minor skews were tolerated in the archived scans. Again, these skews were primarily a reflection of the original prints, although a few were the result of large multiple page imports and bindings.

Migration

Borghoff et al. (2005:XI) refer to the transfer of media type, for example from paper to digital scan as migration, stating that "to date, migration is probably the most common method used to preserve digital data". The authors cite the Task Force on Archiving of Digital Information's definition of migration. "Migration is the periodic transfer of digital materials from one hardware/software configuration to another or from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology" (TFADI, Commission on Preservation and Access, & Research Libraries Group, 1996). The ability to transfer the Sahaptin documents is, therefore, of paramount concern to the permanent archive, but was also crucial for the temporary archive seeing as the formats chosen for the temporary archive needed to provide as much flexibility as possible. The permanent archive might, in time, use a different format for storing data, so the content housed in the temporary archive needed to be easily transferrable as a migration will most likely occur at some later point.

[2] The author would like to thank the University of Washington Language Learning Center for the use of their document scanner for this project.

Even so, further migration is still an issue facing PDFs. Though future formats cannot be predicted, there is always a risk taken when migrating from a compressed, or lossy format, to any other format. If additional compression is added to the file during migration, there will be data loss, and if no additional compression is added, but the same compression functionality is not existing in the new format, there will be an exponential growth in the storage space required for each material. Both of these situations constitute a lossy “transformation” (Borghoff et al., 2005:47). If, however, there is no loss of information and no excessive increase in storage spaced needed, a successful migration is termed a “refresh” (Borghoff et al., 2005:38).

These issues above are examples of why an archive that is intended to remain in a community without an institution that can handle ‘long term’ archiving must expect substantial effort to be required down the road. In addition, while there may not be any intention to transfer the documents to a community-external ‘long term’ archive at the time of the ‘short term’ archives creation, it is vital that the means to migrate it exist as needs and opinions may change in the future, and the community may establish a ‘long term’ archiving institution that could host the content at a later time.

Audio Materials

While the majority of the focus in this chapter regarding the archive deals with the digitization of the paper materials, support for the audio files is required as well. The processes and standards for digitizing have been covered elsewhere thoroughly (e.g. (Plichta & Kornbluh, 2002)) so this section is instead concerned with the process of archiving and the accessibility of uncompressed digital audio files (commonly in either a .WAV or .AIFF format).

Lossless

Working with audio comes with a different set of concerns than digital documents or images. First, in the case of this particular set of materials, the issue of size becomes relevant. The compression rate for the print materials when converted to PDF is fairly good compared to lossless audio files. For archives where text documents are digitized to a higher quality standard, an 8.5” x 11” greyscale 300dpi TIF file totals about 8mb, while a color version would be about 24mb. PDF compression by the XEROX ColorQube 9201 reduces the greyscale file size to under 100k per page.[3]

Comparably, an uncompressed mono 24-bit audio WAV file with a sample rate of 44100hz would be around 465mb per hour of recording time, which, with lossless compression can be reduced to a value around 225mb depending on the content. Recent developments in compression techniques for audio files have resulted in a few options that can substantially reduce the size of a raw audio file (i.e., by some estimations, an average of 40-50%) without any loss of information (i.e., lossless), as opposed to more widely used lossy compression formats (e.g., MP3, AAC). Of these options, it was found that the Free Lossless Audio Codec (FLAC) (Xiph.Org Foundation, 2014b) would be the best fit for the project. While FLAC is not widely supported by Apple OS products, it can be played and edited by many commonly supported open source software options (e.g., Audacity). Apple recently open-sourced their lossless codec and encoder, the Apple Lossless Audio Codec (ALAC) (Apple, 2011), and the compression rates and feature sets between the two

[3] These numbers reflect scans performed with the lowest setting of compression in the scanner’s settings. It should be emphasized that this level of compression is not recommended for any archive. This was a matter of practicality and limitation of the tools at hand. The only benefit to having highly compressed PDFs is for the download time and improved embedding in a webpage. The documents scanned for this archive were primarily computer print outs or multiple generation copies so additional compression was not as much of a concern as it should be for other archives.

formats are very similar. However, for this project, FLAC is preferred as the costs of non-OSX hardware and software are comparably lower and the format is arguably more widely supported.[4] Additionally, ALAC has no built-in error checking functionality, while FLAC does.

Audio Metadata

One advantage many compressed digital audio formats (e.g., FLAC, MP3) have over uncompressed formats (e.g., .WAV, .AIFF) and print (with PDFs) formats is that the encapsulated metadata functionality is much more robust, as they support ID3 (O'Neill, 2013) tagging. With ID3 tags, information on each file, corresponding to the resource' metadata (§0) in the archive, can be added. ID3 tags are widely supported for viewing in audio players and many open-source audio library management platforms (e.g., *iTunes*, *Foobar 2000*) provide options for quick editing at the bulk and individual level. The issue of file accessibility within the temporary archive system will be discussed later on in this chapter (§0).

Video Materials

There were only a few video materials associated with the collection of materials destined for the Sahaptin Language Learning Materials Archive. The issue of archiving video faces many communities and will continue to become more common as time goes on. In this section, I will review some of the special concerns related to video, specifically metadata limitations and problems limiting distribution.

Video Metadata

Extensible Metadata Platform (XMP) (Adobe, 2014b) was originally developed by Adobe for use in Acrobat, but has become one of the more popular standards for video metadata. For the purposes of this and similar archive, the standard is ideal as its codes are based on the Dublin Core Metadata Initiative (W3C, 2008) and XMP was recently adopted as an ISO standard (Gasiorowski-Denis, 2012). Common video formats that support XMP metadata include QuickTime (MOV), Video for Windows (AVI), Windows Media Video (WMV), and MPEG-4 Part 14 (MP4) (Adobe, 2014a). OGG for video does not support XMP, and currently has only limited metadata support using *VorbisComments*. While the more open source nature of the format might be appealing for audio, the metadata and functionality limitations of OGG video make it less than ideal for archiving purposes.

Size and Distribution of Video Resources

It is argued in this chapter that there is a benefit for having a compressed (ideally, lossless) and uncompressed format for a certain file format for instances where accessibility is as much of a concern as preservation. However, the difference in size between a compressed and uncompressed file for audio and images is substantial, but does not compare with video. In fact, providing uncompressed video of any substantial length over a web server is likely prohibitive for most similar archives. For example, thirty minutes of video that has limited lossless compression, 8-bit, 1280x720 with a framerate of 29.97 could be around 138 GB. The same specifications but with the MPEG-4 Part 10, Advanced Video Coding format (aka H.264) default settings would be 13 GB.[5]

[4] As an example, FLAC is the default audio format for the Internet Archive (Internet Archive, 2014). However, as of 2014, FLAC cannot be uploaded directly into the dashboard of the Content Management System selected for this project (Wordpress) without an additional plugin. Manual FTP uploading is required.

[5] All of these numbers are estimates and can vary widely depending on settings and the compressibility of the video (i.e., static shots with minimal changes vs. lots of detail and movement).

Even the compressed file would be expensive to deliver via a webserver. Assuming that the majority of materials for endangered language learning material archives are on lower resolution formats, a thirty minute NTSC uncompressed 640 x 480 file would be around 30 GB. Therefore, in order to provide any reasonable accessibility to the video file, a compressed, and inevitably lossy, file format must be used. Of the formats listed above (§0) that support XMP metadata, AVI and MP4 are arguably the most widely supported by both applications for production and those for consumption. AVI makes a suitable lossless format, and MP4, the compressed deployment copy.

Without substantial available resources, setting up a streaming media server is not feasible. The financial costs alone are likely to be out of the question for most communities. Therefore a compromise must be struck and an external streaming service will likely have to be used. Some communities may have access to a university that would be willing to host the files which could provide some additional control or security. Most communities, however, may have to use an online service such as *YouTube* or *Vimeo* to host their files. The downside to this is that, even with password protected and unlisted files, some control of the files is lost. Regardless, this may be a necessary compromise.

MP4, using the H.264 codec, is a standard supported by most, if not all, major online video streaming services. Google Drive offers similar streaming functionality for video files using the *YouTube* system but keeps the files outside of the YouTube system proper, although at the time when this chapter was written, there are significant issues with the Google Drive video streaming system that make it impossible to recommend.[6] Once files are hosted, and set as unindexed for searching, in a streaming service they can be embedded within the archive CMS on the proper resource page with the metadata. At a bare minimum, the video's page on the service (e.g., YouTube page) should contain a link to the URI/URL for the resource in the archive.

Backups of Uncompressed Video

Finally, there is the issue of storing the uncompressed files. It is possible that some communities could have terabytes worth of files. If resources are available, a service offering low cost, infrequent access, online storage could be used (e.g., Amazon Glacier). A less ideal, but more cost effective solution for community archives that don't have the resources for long-term archiving is the following. First, multiple PC workstations at multiple locations will be needed, as it's vital to have offsite backups in any situation. Each of these workstations should be equipped with hard disks[7] dedicated for the archive. Each workstation would then install *BitTorrent Sync*, or a similar service, that will securely keep the data consistent on all of the workstations. The service transfers files in secure packets from one computer to any other peers (i.e., the other workstations) until all of the data on the transferring drive or folder matches the peers. With such a service, there is no reliance on the cloud, and so data security is improved and cost is limited to the purchase of any physical hard drives and workstations.

[6] For a test, some video files were deployed using Google Drive and embedded them elsewhere. All video files were played the same amount of times, but seemingly at random, certain video files were blocked by the system for further streaming. This issue could only be resolved by deleting, re-uploading, and re-embedding the files.

[7] A reviewer recommended enterprise class hard disks with a predicted high Mean Time Between Failures (MTBF) number. Organizations can also increase the life of the hard drives by limiting negative environmental factors, such as excessive heat or vibration. However, no assumptions about the life of the hard drives should affect the implementation of a redundant, live, offsite backup as described above. As long as there are multiple computers which as properly set up to handle the backups, multiple consumer grade hard drives are much better than one or two enterprise grade drives.

2 Ethics & Security

As opposed to many cases of the digital archiving of Indigenous materials (e.g., (Christen, 2011)), the digitized versions of the Sahaptin learning materials are intended to replace the originals in most cases. The original documents were collected by Edward and Mary James for the sake of archiving and had previously been inaccessible to educators and learners, so this, new digital archive brings an enormous increase in accessibility. Documents that once occupied 20+ banker boxes are now filtered for redundancy and searchable using a variety of means.

Copyright

The issue of copyright is connected to a few dilemmas facing the archive, particularly access with respect to ownership and National/Tribal/Community preferences.

> *"The heart of the problem lays in the intent of copyright. Copyright is meant to protect individuals, creators, while ensuring the public has a level of access rights. This is a (sic) idea meant to promote propagation while individuals are compensated for the dispersal. In this situation, the net effect is that a communally owned creation is considered public domain, or free for all. Communal creations have no individual to protect; what is more, since there is no individual, everyone owns it" (Wolf, 2008:138).*

Francis & Liew (2009) echo this concern that copyright is oriented towards the individual and does not account for Indigenous concerns, but notes that efforts are underway to change this. Another issue they raise is that definitions of legality may differ between nations, with the dominant (i.e., usually non-Indigenous) population's definition and needs overruling those of the minority nations. Christen (2008) also, deals with the complexities of balancing security and accessibility.

The majority of the content in the Sahaptin Archive was created by Dr. Virginia Beavert and Edward James, along with Mary James who donated the materials to be archived and made available in the first place. However, some items contain content that may not have been created by Beavert or James and, instead, belong to, as an example, the cultural heritage of the Yakama nation. Although these materials were allegedly used in a classroom and, as such, are likely not as potentially culturally volatile as pure documentation or text collections (e.g., (Innes, 2010)), there may still be content that is considered inappropriate to share with an outside community. While some form of security for sensitive content is necessary, the more difficult issue arises around how to determine what content should be secured. The protocols for Native American Archival materials state the following: "Consult with culturally affiliated community representatives to identify those materials that are culturally sensitive and develop procedures for access to and use of those materials" (First Archivists Circle, 2006), so the security used in the Sahaptin Archive requires the oversight and curation of community members. Various members of the Yakama Tribal and General councils were notified about the projects by phone with multiple voicemails and email when available.

Security

Security issues for the archive were a crucial topic of concern and were included in a needs analysis undertaken to support this project. (Christen, 2011) reminds us that the structure of community involvement and oversight for the archive need to be resolved before an archive can be considered secure. Should the needs analysis find approval of the proposed password security structure of the Sahaptin Archive, one must then ask who should be in charge of distributing and maintaining the passwords and if community members should be able to annotate and curate the materials in the archive versus the archive remaining static with the sole control being related to access. The temporary archive conforms to the Open Archival Information System (OAIS) (CCSDS, 2012) whereby a primary goal is to preserve the materials for the ‘long-term’. The OAIS standards describe ‘long-term’ as potentially indefinite, but the Content Management System (CMS) deployed version of the archive is not a feasible long-term solution in the case of the Sahaptin material for a variety of reasons, including the lack of a dedicated, long-term archivist. Therefore, the permanent archive will need to be responsible for the indefinite storage of the materials in a way that is fitting considering the limited parameters, and the OAIS standard of long-term should apply to the CMS archive in the sense that it should be made as perpetual as possible.

Even though the word “open” appears in the standard’s title, restriction and security are still a part of the system. The OAIS prioritizes authenticity and consistency or accuracy of the data. Therefore, removal or editing of the core documents goes against the standard. An OAIS archive must ensure that information “is never deleted unless allowed as part of an approved strategy. There should be no ad-hoc deletions.” (CCSDS, 2012, s.3 p.1) However, editing or management of security and the metadata, as well as annotations or corrections, are accepted provided producers (i.e., editors) follow the prescribed submission agreement (CCSDS, 2012, s.2 p.9). The submission agreement, in the case of the Sahaptin Archive, clearly outlines the types of edits and additions to the accompanying CMS-based (i.e., not the PDFs) content (e.g., metadata, annotations) and can be accessed by anyone given the proper permissions. While the system developed for this temporary archive attempts to adhere to the processes laid out by the OAIS, many of the processes have been combined and simplified where appropriate for efficiency and due to the practicalities of the tools available.

3 Methodology

Labelling

Bird & Simons (2003) describe the seven pillars of archiving as “content, format, discovery, access, citation, preservation, and rights” while Johnson (2004) argues for an eighth pillar, labeling, saying, “Nothing could possibly be more important than labelling every single item you produce—each track, tape, disc, notebook, digital file, photograph—with RUTHLESS CONSISTENCY” (p. 8). As such, the labelling system for the Sahaptin Archive now exists at four primary levels:

1) Filenames, which use the following conventions:

   EXAMPLE: S_C_0000-Example_Title.pdf
   GLOSS: 1_2_ 3 - 4.pdf

- 1: (S = Sahaptin).[8] While largely redundant, this name is intended to anchor each file within the larger collection.
- 2: This position will contain a character or two relating to the primary category of the material. (e.g., C = Culture, S = Songs, P = Prayers, etc.)
- 3: This position is an arbitrary four digit hexadecimal code used to link the material to the online database.

[8] The use of a three letter ISO language code was decided against to minimize the filename length, because the archive will only contain a single language, and because the Linguist List standard (qot) will likely not improve the clarity for the majority of users and future administrators. However, the ISO code for the dialect (ISO 639-3: yak) was used in the OLAC metadata tags where appropriate.

- 4: This position is for a short name used to roughly describe the material.

  The entire filename serves as a unique identifier in the system and XML. The filename should be located within the following XML tag set:

  ```
  <dcterms:identifier xsi:type="dcterms:URI">Example
  Identifier</dcterms:identifier>
  ```

2) Folders, which serve as a redundant organization of the materials within their larger categories. In addition to metadata tagged resource files being stored in the online CMS, all files are stored on the development computer and backed up on various servers. Outside of the CMS, all files are stored in a file folder structure corresponding to the larger categories and noted in the Google Apps Spreadsheet.

3) Online labelling, which provides ease of access. There was little consistency to the original organization of the materials, so the slightly arbitrary labelling of the filenames is the result of the variety and complexity of the materials delivered to the UW. The bulk of the labelling, therefore, reside in the online archive CMS in order to take advantage of a combination of single grouping labels (categories), overlapping labels (tags), and additional metadata related to the PDF (e.g., the hexadecimal code linked to the file name).[9] The metadata may overlap with the tags or categories assigned to each resource in the CMS, which will be discussed later on. The *bundling* (Johnson, 2004, p.9) of the various related materials was done using tags in order to allow multiple bundles of the same resources, which mimics much of the original boxed organization and allows the static bundling of individual resource pages for more robust bundles.

4) The Uniform Resource Identifier (URI), which identifies each archived resource and will likely be altered when the resource enters a permanent archive. Steps were taken to follow URIs & Uniform Resource Names (URN) specifications (Borghoff et al., 2005) and Linked Data standards were adhered to as closely as possible. The Filenames described in (1), above, account for the URN. The URN is then assigned to the file and added to the online material database (Google Drive Spreadsheet) and metadata and the CMS site. The CMS that the temporary archive is using is Wordpress, which is a widely supported and accessible CMS. Wordpress automatically generates a Uniform Resource Locator (URL) for each post associated with a resource, though the URL (i.e., also "permalink" in Wordpress parlance) could also be manually specified.
   Thus, instead of having an auto-generated URL using the title of a resource (e.g., http://depts.washington.edu/sahteach/resource-title/) the URN could be substituted for the title or for any other default URL pattern, which would result in something like the following:
   http://depts.washington.edu/sahteach/S_T-0056-Handout_Mish_-Kan/
   At this point, the URIs in the temporary archive now (depending on the structure of the permanent archive) have a slight advantage. A permanent archive could, should they deem it worthwhile, fairly easily reconstruct the structure of the temporary archive (minimally "disrupting" the URI/URL pattern) by swapping out the URL content prior to the URN (e.g., http://depts.washington.edu/sahteach/). On the other hand, if the default URL setup is used where the title of the resource is used to generate the full

[9] The OAIS would define this as an identifier related to 'Reference' as part of the "Preservation Description Information."

> URL for the post, the title is included in the metadata and, barring any special character conflicts, a script could be written to keep the URIs more consistent. In the end, the extensive metadata, file naming and Google Spreadsheet info may be more than enough for a permanent archive to efficiently intake the content, and added redundancy will be no more than a few added clicks.[10]

Core System

The types of materials contained in the archive differ in format from the content usually found in major online language documentation archives. Crucially, the Sahaptin materials lack the naturally produced language (i.e., texts) or pure linguistic documentation (e.g., grammars and lexicons) often seen in major archives. The materials were designed, collected and organized with the single goal of pedagogy in mind, so a stand-alone archive was selected to better accommodate the format and genre of the content. Based on a brief anecdotal survey, similar community-based archival efforts have much in common with the Sahaptin archive such that using an isolated stand-alone archive that could be transferred to a permanent archive at a later time offers advantages for production, management and make the archived materials available to community members sooner.

To conform to the OAIS, a concerted effort was made to develop an archive and management system that was "Independently understandable" (CCSDS, 2012, s.3 p.1) in regards to access or being 'designated' to the community. It is vital the materials deemed appropriate for distribution (§2) are easily accessible to the community and that the stand-alone option allow for the navigation and presentation of more colloquial, less technical (e.g., linguistic) jargon. Ideally, the successful accessibility aspects of the short-term archive will be present in the permanent archive as well.

For this project, the Wordpress Content Management System (CMS) was selected, though other CMSs offering expanded functionality might be worth investigating for other projects. For example, Drupal's extensive user and permission system could be useful for more extensive archiving projects or for projects with more people working on system administration and content submission/curation. Wordpress, though, met the less extensive requirements of the Sahaptin Learning Materials archive and, as such, would be a good fit for other similar archives.

As of 2014, Wordpress was the most frequently used CMS in the world (Builtwith.com, 2014), and this widespread usage correlates into substantial support options, including documentation and forums, and arguably requires less technical expertise to install and manage than some other CMSs (e.g., Drupal, Joomla). Wordpress is easily updated through the in-browser administrative interface and can be password protected at the site and page/post levels, though multiple users accessing pages for editing purposes will require some form of rollback (e.g., Wordpress *revisions*).

When it comes to features related to archive distribution, Wordpress supports the tagging and categorization of posts. Since each resource is associated with an individual post, the tags and categories are easily assigned to the resource, and they are then included in the search function, and may be explored and displayed in a variety of ways.

Future migration of the archive was also considered when Wordpress was selected as the CMS for this archive. Wordpress uses MySQL for its database, so when it is transferred to a permanent archive the full data (e.g., the pdfs, texts, tags, etc.) could then be extracted and, with a bit of scripting, relatively quickly transferred into the new archive. However, it is important to note that if automated backups are not performed on the server side, a backup plugin should be installed (e.g., UpdraftPlus) to regularly create offsite backups of the system and archive content.

[10] Manually altering the permalink requires two additional clicks and one copy and paste of the URN during the post building process, which is discussed later.

Some additional plugins were also be of use to this and may be valuable for other archives, including those that do the following: index PDFs for searching (e.g., *SearchWP*); truncate posts in the search results (e.g., *Post Teaser*); embed and display the PDFs inside each post (e.g., *Spiderpowa Embed PDF*); automatically format the XML metadata (e.g., *SyntaxHighlighter Evolved*); and embed audio for streaming or streaming-like behavior (e.g., *HTML5 jQuery Audio*).

User Interface, Design & Accessibility[11]

Another topic that deserves consideration is accessibility as it relates to user interface (UI) and aesthetics. Archives can have an enormous set of complicated data, not all of which is very friendly to human readers. In order to make the short term archive as usable as possible it was important to make it relatively aesthetically pleasing to visitors so that they would not be overwhelmed, or simply bored, by the content. Also, the content should be packaged in a way that shows respect to the creators of the content and the community it represents. For example, a fairly easy solution was to purchase a single high resolution image of some nature near the Yakama community from a reputable stock photography company and use it as an entry header and site footer.

More obvious UI modifications to the CMS were required as well. Once a base Wordpress theme has been chosen, a child theme must be immediately created so that future updates will not destroy any modifications made up to that point. After the child theme was created, all of the basic functionality and auto-generated content (e.g., calendars) were removed excluding a limited core set (e.g., menu bar, search box). The content that was featured on the main page was limited to some brief information about the archive, a link to a page with all of the categories for the content, links to pages on ‘How to use this archive’ and ‘The History of the archive’, an auto-updated list of recent additions to the archive (to inform as well as highlight growth), and below the initial screen a set of links to related resources. The ‘How to use this archive’ page provides detailed explanations for people trying to find content (via labelling (§0)), as well as guides on adding or transferring content. Some redundancy was added to the interface, such that the header and footer on every page provided access to the tag and category links. Since the needs analysis resulted in no interest in having mobile support for the archive and because the archive will eventually find a permanent home, time was not invested in making the theme fully responsive. Finally, every theme tested for the archive required substantial modification of the CSS for the fonts. A special set of CSS styles needed to be constructed to display certain Sahaptin characters correctly. Again, because it is easy to overload the user with too much content, it is important to utilize font sizing, spacing and padding appropriately. A well-constructed child theme with robust CSS can provide additional longevity and portability of the archive.

Access controls & Security

Because some of the content in the archive may be information the community wishes to protect, a form of access control needs to be provided. Christen discussed this issue in her work on the Mukurtu Wumpurrarni-kari Archive, stating that the needs of the community as well as the sensitivity and complexity of the archive necessitated a tiered, or granular, permission structure (Christen, 2011). Some Mukurtu Wumparrarni-kari materials, for instance, needed to be restricted such that they would only be accessible by women in the community. In the Sahaptin Archive, the limited content type and concerns over long term maintenance led us to choose Wordpress as the

[11] The modified theme and XML postbuilder web application code have been posted on the archive’s website for those interested in starting a similar archive to take and modify as needed. http://depts.washington.edu/sahteach/introduction/how-to-use-the-archive/#start

archive's CMS because Wordpress allowed two permission tiers. Should site-wide password protection be needed at some point, the entire installation could be protected with a .htaccess file, although this might require the development of an additional public intro page describing the archive and access information. The password protection of individual posts and of the site as a whole could then require some degree of moderation or administration, especially in the provision of access to new users or in instances of password change if the site became compromised. For the sake of local control, engaged community members would be ideal for this role. Johnson provides another perspective on the issue of preservation versus accessibility, stating, "Archived materials are public goods, even when access is restricted to protect the rights or wishes of the speakers whose words are recorded therein" (Johnson, 2004:3). Again, communities where multiple editing accounts are required should consider the functionality options of other CMSs (e.g., Drupal) as well.

While restricting access to certain content is important, nearly all digitized content, once made accessible, can find its way outside of the control of the archive as "true security would entail not allowing one's data on the Internet" or on any computer connected to the internet (Wolf, 2008:140). Wolf argues that security for Indigenous resources hosted online is an imperative, recommending precautions such as geographical (IP-based) restriction, SSL encryption, etc. However, security must be balanced with accessibility, something Wolf addresses as well. Another concern is that some communities may need to consider the costs involved in options like SSL, whether financial or require expertise.

For this reason, it is often best to utilize techniques that could slow down or frustrate pirates. For example, each resource in the Sahaptin Archive was given its own page (or, 'post' in Wordpress parlance) making it much less convenient for someone to program a script that could spider the page and automatically download each PDF. The PDFs were then embedded for display in the browser, another function which should discourage unauthorized downloading.

It is possible to protect things at the file level outside of Wordpress as well, so an added benefit of the PDF format is that proprietary software (e.g., Adobe Acrobat) can be used to restrict access with encryption. Currently, for the PDF format using Adobe Acrobat there are three options for encryption available: Password encryption, Public Key Infrastructure (PKI) encryption, and Rights Management. The last of these might be a useful option for Indigenous organizations wishing to restrict access to a particular physical environment, on say, an internal network.

Determining which Sahaptin resources contained sensitive information went beyond my expertise, as I am not a member of the Yakama community. And it is important to remember that even within a given community, there is always a chance of disagreement or gradual attitude change over time. For this reason, the Sahaptin Archive was initially developed with an option for members of the community to request that content become protected with a password so any documents containing potentially sensitive cultural material in the pilot phase could be password protected as a default. Additional guidance on the archive and security was a key goal of the needs analysis that was conducted for this project.

In the end, it is important to carefully weigh the situation before placing strong restrictions on content. Even if content is restricted to a few people in a community, it is necessary that it be as accessible as possible to those people. Edmondson's report for UNESCO makes the following point:

"Preservation and access are two sides of the same coin. For convenience they are considered separately in the following discussion, but they are so interdependent that access can be seen as an integral part of preservation. Indeed, the widest definition of preservation embraces almost the totality of an archive's curatorial functions…Preservation is necessary to ensure permanent accessibility; yet preservation is not an end in itself. Without the objective of access it has no point. Both terms have a wide spectrum of meaning, however, and tend to mean different things to professionals in different situations" (Edmondson & UNESCO, 2004:19) .

Embedding and Streaming Audio

Returning to the issue of audio files in the temporary archive, it would be ideal to have the master archival file also be streamable or playable in a web browser, similar to the method of embedding the PDF discussed above. Thus, in a perfect world, the audio master that is encoded as a FLAC and tagged would also be playable in a browser. One recommendation for this is to encapsulate a FLAC encoded file into an Ogg container (Xiph.Org, 2013; Xiph.Org Foundation, 2014a). At the time of writing this chapter, however, this technique is not widely supported and Wordpress does not offer any stable plugins for this purpose. For most communities, this option may not be feasible depending on the resources available. This leaves two primary remaining options. The first is to simply require the user to download the FLAC before listening to it. The second option is to encode a second file in the lossy MP3 or Ogg Vorbis file format and embed it into the Wordpress site (using a plugin like *HTML5 jQuery Audio*) accompanying the downloadable FLAC file. Crucially, the lossy files must be kept separate from the actual archive data structure. Although this adds a few additional steps and clutters the temporary archive it has an additional advantage in that the reduced size of the lossy file will be more accessible to communities with poor or unreliable internet access. While it is dependent on the bit rate at which the lossy file is encoded, the reduction in file size from a FLAC to an MP3 could be more than 50%.

Collaborative Editing for Audio Transcripts and Error Corrections

While a transcript can be easily added so that it accompanies its associated audio resource in an online system like the one being described in this chapter, the issue of facilitating collaboration on such a system is not so straightforward. Wordpress has multiple user roles but for the collaborative functionality that might be required for documents like transcripts there are only two relevant roles: *Administrator* and *Editor*.[12] Both roles have permission to edit any existing post, with administrators having access to additional system options. This means that an editing user cannot be granted permission to edit a single page/post, or a section of one, they must either have access to all of the content in the site or only have the option to create new pages. For smaller communities where there will only be a few trusted people editing the site this should not be a problem. However, if a more robust set of user permissions is required, another CMS, such as Drupal might be preferable. Wordpress does, by default, retain the last 25 revisions to any page in the site, so unless a user takes the time to maliciously edit a page 25 times and no recent back up was made, there should be very little risk of data loss using this method.

Another way to address collaboration could be to utilize a wiki. For example, an installation of MediaWiki could be deployed solely for these purposes, while for smaller archives, Google Drive could be used. Each resource entry URL/URI in the Wordpress archive could have a link to a page in the wiki where the transcript would reside. Or, after a transcript that has been edited in the wiki is deemed suitable for the archive, the text could be migrated to the Wordpress site for visitors to access. The benefit of this method would be to limit access to editing the core Wordpress data. On the other hand, this would add some unnecessary disconnect and spread of the data, which could add unwanted complication when data is migrated to a permanent archive as there would be two databases (on for the Wiki and one for the Wordpress installation) that would need to be joined. There is also a risk of content being added in the wrong location or extra redundancy. In addition, locating transcripts in an external system limits the power of the CMS search engine.[13] If a

[12] Users assigned any of the other editing roles can only edit posts the same user created.

[13] It is not a simple task to setup a Wordpress search box which has indexed the content of a separate MediaWiki installation as well as the Wordpress content. Even if it was setup properly, either the user would be directed to the

transcript accompanies the URL/URI for the resource in the CMS then each word will be indexed. Hypothetically, extremely extensive data could clutter the search database, but with tags and categories as alternative routes to resources, this is not as much of a concern for the type of archive in question. There are some plugins available for Wordpress that provide wiki-like functionality within the system, but, at the time of writing this chapter, none of them integrate the content into existing posts, and the only plugin with reasonable support and positive reviews would be cost-prohibitive for most communities.

One of the benefits of existing collaborative software options is the ability for users to comment on the content and edits. If this is crucial functionality for a community, external systems could be used or a plugin for this purpose may be developed at that time. In the end, the existing revision functionality in Wordpress was judged to be preferable for this archive as it allows all of the content for a resource to be located under a single URL/URI and it reduces the amount of software, plugin or server support and maintenance that will be required.

Labelling relevant to the Indigenous community but potentially outside of a standard metadata schema also can be added to the resource's respective post (Toner, 2003). As discussed earlier, the original learning materials sometimes contained errors or required additional notes. If the notes and corrections are reasonably few they can be added to the metadata and description for the resource, but if extensive they should be listed in section for corrections above the metadata block on the post/page. Although the long-term feasibility is questionable, an additional step to improve the usability of the resource would be to add corrections to the PDF itself using the comment and text box functionality. This method, however, does severely hinder collaboration as the edited PDF must be re-uploaded to the Wordpress system and the URL to the file updated in the metadata if it is not replaced. Because there is no online logging for the edits or check-in/check-out functionality this method should probably be considered only as something to be done in addition to the corrections text accompanying the metadata on the Wordpress post. It should be noted that if errors are plentiful for a particular resource, barring a substantial reduction of quality, the original resource should be corrected in the PDF and then added to the post and metadata with the same URI/filename with some additional text in the filename marking it as corrected (e.g., '_updated').

At this point, there is still the issue of how to display corrections if they are disconnected from the actual image of the resource (i.e., in the PDF). For example, one option would be to list the page number and the line where the error is found, write the error, and then have the correction or corrected text immediately following it. If possible, it is ideal to have some distinguishable string of Unicode text that can be found via search, within the browser for the post and within the PDF viewer for the actual resource. This refers to situations where a string of Sahaptin that is not either not search friendly in the browser or OCR was not successfully ran on the PDF. If this string of text is located within a resource of 50 pages and 100 lines per page, finding the location of the error without the page and line number location could be incredibly cumbersome. Thus, if the incorrect text could be located via a search that would be ideal. If the string immediately containing the error is not searchable, then a neighboring chunk of text that is searchable, and as unique as possible,[14] should be included in the 'error string' on the resource page.

The final set up implemented for the online archive was the addition of three presentational boxes for audio transcriptions, "Retyped" (PDF to Unicode) transcriptions, and annotations. Each box type is color coded and located immediately below the display for the resource (e.g., the embedded PDF). The Unicode transcription box is a place where text on a PDF that isn't machine

MediaWiki transcript page, thus leaving the system and having incomplete info on the resource, or it would direct them to the Wordpress resource page where the transcript text from the search that prompted them to click the link would be absent, as it is stored on the MediaWiki page, separate from the primary resource page in Wordpress.

[14] E.g., the two word chunk 'velvet cake' is likely more unique than 'it is'.

readable can be posted in a machine readable (i.e., Unicode) form. If a document has had partially successful optical character recognition ran on it then that text could be corrected in a text editor and added to the resource page. If a document was hand written it requires a full manual transcription. While corrections and annotations should be noted in a correction/annotation box, ideally, the text in this box should incorporate the corrected and properly annotated text (i.e., be the most accurate version of the document possible). While the majority of Sahaptin characters will render properly in a text-editor and HTML, *X̲* will not render properly in HTML when copied from a word processor. Instead it will generate the underline on the following space or character. However, if you have a standard Sahaptin keyboard and font installed you can copy and paste the text back into a word processor and the underline should display correctly again under the *X*.

In the end, these decisions related to annotations and curation are for each community to make depending on the types of resources and errors involved, although the above method seems to be a suitable solution to the problem for the Sahaptin archive.

Long Term Archiving

The recent CCSDS recommendations document for OAIS (CCSDS, 2012) describes some of the current challenges for 'long-term' archives. The most crucial of these considerations are data corruption, hardware failure and hardware obsolescence. As all of these will likely be of concern to the permanent archive, regular backups of the MySQL data will be made on dedicated servers hosted by the UW to better compensate for these issues on the CMS hosted short term archive. However, it should be emphasized again that these problems will arise eventually for any community that wishes to internally manage an archive without the support of a 'permanent' archival institution, whether external or internal to the community, underscoring the definition 'short term'.

In order to better account for these risks and problems all documents in the Sahaptin archive were labeled and organized into folders, along with a dump of the launched archive, and packaged into a single compressed file to be: 1) kept by the online archive builder; 2) put into some form of cloud storage (e.g., Google Drive Spreadsheet); and 3) shared with interested related institutions (e.g., The Yakama Nation, The Northwest Indian Language Institute (NILI), Heritage University). Regular XML exports of the metadata and annotations, if updates are made, will then be easy to provide to the permanent archive. Other community-based, or isolated, 'short term' archives should follow the steps above or something similar.

Metadata

The temporary online archive is primarily based on Dublin Core Metadata Initiative (DCMI) (DCMI, 2014) standards, which are commonly used for similar archives (e.g.,(Christen, 2011)). In addition, any data points not covered by Dublin Core are covered by Open Language Archive Community (OLAC) standards (http://www.language-archives.org/documents.html#Standards).

Borghoff et al. (2005:50) argue "robustness must take priority over efficiency" with respect to metadata projects destined for long-term archiving. However, in this case, an extreme reduction in efficiency (e.g., adding many more steps to the process) would have slowed down or prevented the completion of the archive. Theodore Gerontakos, the XML specialist for the UW libraries, worked on the initial set of proposed metadata tags to capture the materials' core information without cluttering the database with tags marked unknown.

The XML and Metadata management application

Each resource or item in the archive needed to have accompanying information (metadata) in Extensible Markup Language (XML) in order to transmit information to current and future users and to distribution systems/archives as effectively as possible. To a human reader, XML is less felicitous. Editing XML manually is a time consuming process and the potential for errors is high.

After the initial metadata XML structure was developed, Gerontakos recommended some metadata completion software options, which guide the user through the fields and limit errors, but the effective software was priced far beyond the means of this project and is likely also for many communities with limited resources. Therefore, it became necessary to develop a PHP-based online application that provided the necessary functionality for a project of this more limited scope. The first version of the application provided fields for the user to complete on the left-hand side of the screen. When the necessary fields were completed, the user clicked submit and the system exported the metadata with the correct info. Unless there were major errors in the input fields, every export fully met the W3C XML standards. This code could then be copied and pasted into the appropriate Wordpress post field. After completing the system, the entire process was reviewed by Gerontakos for potential errors. Soon after, it became clear that with a minimal expansion of the fields and scripts the system could handle the entire code set (e.g., title, PDF link, embedded PDF code, general formatting) for each Wordpress post. The XML application was then reworked into the 'Post Builder', reusing some metadata input fields and adding in others. The final application outputs a single chunk of code such that no edits were required in the Wordpress post's description box after the generated code was posted within, thereby reducing the risk of error and the number of steps required to add a single resource.[15]

Most of the code generated will never need to be altered (e.g., the Dublin Core and OLAC declarations), though the input fields in the generator represent the XML tags that are most likely to be altered for the purposes of this archive and are arranged in the predicted frequency for which they will be edited (i.e., top=more, bottom=less). Any tags that are not represented with a field may also be edited manually after the generation process. The system is a mix of PHP, HTML and CSS and is intended for portability and ease of modification. The entire application, excluding the help page, is contained in a single PHP file.

Using text resources as an example, the final workflow for the archive is as follows:

1) Open an unarchived PDF.
2) In the Google Drive Spreadsheet,[16] create a new name (URN) for the resource.
3) Change the name of the PDF and move it to the appropriate folder.
4) Upload the PDF to Wordpress.[17]
5) Copy the URL of the PDF stored in the Wordpress installation.
6) Paste the URL and URN into the Post Builder system.
7) Complete all of the fields that are relevant to the resource. Most likely, the user of the Post Builder system will never need to edit the Linguistic field, Subject or Type. These are set to default values and may be left as is unless they need to be changed. Complete at least Title, Date and Description and then as many other fields that are relevant to the resource.

[15] PHP, unfortunately, does not have a quick way of sending echoed text, or text in a div, to the clipboard. HTML and Javascript options are not well supported cross-browser. Flash based clipboard scripts also have support and security issues. Thus, for the sake of limited maintenance and better security, the selecting and copying must be done manually. The text of the exported XML is generated especially small in order to make the process of selecting and copying easier and quicker. If it necessary to examine the generated code, it can be pasted into a text editor or Wordpress.

[16] This spreadsheet is a redundant knowledge base to stream line the administrative process. It will not be required for the archive in the long term.

17 E.g., (OAIS) archival storage.

8) Click Submit.
9) After the code is generated,[18] select all of the text and copy it (right-click 'copy' or ctrl-c).
10) Paste it into the Wordpress (the main archive site) Post field at the bottom of the entry.
11) Enter the title of the resource for the Wordpress post.
12) Edit the permalink, paste the URN, and click *OK*. (This step creates the URI)
13) Add all appropriate tags and set the category.
14) Click Update.

Post Builder Input Field Descriptions

Below are simplified descriptions of the fields and how they relate to the metadata.

- **Title:** The name of the resource.
- **Date:** The date the resource was created. If the resource was created in 1995 but incorporates another resource (say, a text from 1918), use the most recent creation date in this field and then enter the older date in the Additional Date field. Acceptable formats for the date can be found here: http://www.w3.org/TR/NOTE-datetime
- **Description:** Any additional info about the resource. For example, a sub-title might go here.
- **Resource URL:** This is the URL for the file once it has been added to the online (Wordpress) system.
- **Resource Filename:** This is the URN, or filename without the filename extension. For example:

  ```
  <dcterms:identifier xsi:type="dcterms:URI">S_L-0059-
  Early_Dictionary</dcterms:identifier>
  ```

- **Type:** The type of physical resource. Is it a text (paper, book), sound (audio recording), etc.?
- **Linguistic Type:** The type of resource linguistically, primarily related to language documentation. See here for explanations of each of the options: http://www.language-archives.org/REC/type.html
- **Format:** The digital format of the resource. See for more info: http://dublincore.org/documents/2012/06/14/dcmi-terms/?v=elements#format
- **Table of Contents:** If the resource has sub-units, such as chapters, they are listed here.
- **Linguistic Field:** This is the sub-field that is associated with the resource. Most likely it will be applied_linguistics. See here for more info: http://www.language-archives.org/REC/field.html
- **Subject:** The subject of the resource, most likely 'Teaching the Sahaptin/Yakama Language'
- **Publisher:** If no publisher exists, leave blank.
- **Speaker name:** For audio recordings and transcribed speech, enter the speaker's name here.
- **Extra Depositor Name:** If you are not one of the original workers on the project (Beavert, James, Hargus, etc) enter your full name here.

[18] This code could be considered as a Submission Information Package (SIP) following (OAIS) standards (Borghoff, Rodig, Scheffczyk, & Schmitz, 2005).

- **Additional Date:** see (Date) above.
- **Restricted Access:** Check this box if the resource has sensitive content and should be restricted such that only members of the Yakama Nation can access it. Checking this box does not secure the content. It only informs future archives that it should be secured. Any current archive systems will have to restrict the content in their own way.

Additional Metadata Info

The system also needs to declare in the metadata the *subject* language of the resource and the language the resource is written *in*. If the resource is written *about* the language, the following tag is used:

```
<dc:subject xsi:type="olac:language" olac:code="yak" />
```

If the resource is written *in* the language (e.g., Yakama), the following tag is used:

```
<dc:language xsi:type="olac:language" olac:code="yak"/>
```

OLAC argues for the use of the *contributor* tag in place of a *creator* tag as well, claiming, the "Recommended best practice is to use the Contributor element instead of Creator, except in cases where there is significant creative involvement by the person or organization and there is no suitable refinement term from the *olac:role* scheme to use with Contributor." http://www.language-archives.org/NOTE/usage.html#Contributor[19]

Because authorship of all of the materials is unknown, the rights holder tag was omitted, but Dr. Beavert is listed as the compiler of the materials. Another potentially gray area was the code for the *linguistic-field* tag, for which OLAC provides the following clarification: "Language Acquisition may be used to describe materials relating to either adult or child language acquisition, and to either first or later language acquisition. However, if the materials deal specifically with language teaching, or with the process of language learning from a pedagogical point of view, they may be best classified as Applied Linguistics." http://www.language-archives.org/REC/field.html

4 Conclusion

Other recent archival efforts related to endangered language materials have been described in the literature. The Dena'ina Language Archive is a recent digitization and archiving effort utilizing metadata standards is backed up in the Arctic Region Supercomputing Center and available via an online interface (Holton, Berez, & Williams, 2007). While approximately a third of the size of the Sahaptin Learning Materials archive in volume (Alaska Native Language Archive, 2011), the contents of the Dena'ina Language Archive are considerably more diverse and complicated. The archive also utilizes a higher quality file format (TIFF), which was not feasible for the Sahaptin Archive as described in this chapter (§0).

This chapter, therefore, describes a single case of constructing a short-term archive as a staging platform for a long-term or permanent archive. Depending on the flexibility of the future archive, the short-term archive may exist after the materials have been transferred if the short-term archive continues to provide an advantage to the respective community (e.g., better accessibility).

[19] "Generally, copyright protection extends to two elements in a sound recording: (1) the contribution of the performer(s) whose performance is captured and (2) the contribution of the person or persons responsible for capturing and processing the sounds to make the final recording… Under the 1976 Copyright Act, which became effective January 1, 1978, a work is automatically protected by copyright when it is created" (United States Copyright Office, 2014:1).

For communities interested in a similar archive for their materials, the system above is one possibility, though it is important to remember software and other options will change and that the community may need a more robust permission structure that other systems offer (i.e., similar to that of the CMS Drupal). However, the pilot of this system has been very promising and response by the community has been positive. If a community decides to pursue a similar archive and wishes to keep materials out of a permanent archive for control reasons, a more extensive backup system will need to be implemented and much more careful migration planning undergone, as well.

While the metadata constructed by the Post Builder is not as complete as they could be, they do cover the key elements that the materials for this archive needed. There is also, undoubtedly, an advantage to keeping the database as clean and as well organized as possible in anticipation for the transfer to a permanent archive. After the post builder application was integrated, including the metadata into the process required few additional steps. Thus, it is completely feasible for any community to include metadata and follow the core Linked Data standard and OAIS protocols. With a basic understanding of HTML and PHP, the post builder application can be modified to suit the needs of any similarly scaled short-term archive.

References

Adobe. (2014a). After Effects Help / XMP metadata. Retrieved from https://helpx.adobe.com/after-effects/using/xmp-metadata.html

Adobe. (2014b). Extensible Metadata Platform (XMP). Retrieved from http://www.adobe.com/products/xmp.html

Alaska Native Language Archive. (2011). Dena’ina. Retrieved October 15, 2014, from http://www.uaf.edu/anla/collections/denaina/

Apple. (2011). Welcome to the Apple Lossless Audio Codec Project. Retrieved from http://alac.macosforge.org/

Bird, S., & Simons, G. (2003). Seven Dimensions of Portability for Language Documentation and Description. *lan Language*, *79*(3), 557–582.

Borghoff, U. M., Rodig, P., Scheffczyk, J., & Schmitz, L. (2005). *Long term preservation of digital documents*. Berlin; New York: Springer-Verlag. Retrieved from http://dx.doi.org/10.1007/978-3-540-33640-2

Builtwith.com. (2014, October 13). CMS technologies Web Usage Statistics. Retrieved October 15, 2014, from http://trends.builtwith.com/cms

CCSDS. (2012). Reference Model for an Open Archival Information System (OAIS). The Consultative Committee for Space Data Systems. Retrieved from http://public.ccsds.org/publications/archive/650x0m2.pdf

Christen, K. (2008). Archival Challenges and Digital Solutions in Aboriginal Australia. *SAA Archaeological Recorder*, *8*(2), 21–24.

Christen, K. (2011). Opening Archives: Respectful Repatriation. *AMERICAN ARCHIVIST*, *74*(1), 185–210.

Cushion, S. (2004). Increasing Accessibility by Pooling Digital Resources. *ReCALL*, *16*(1), 41–50.

DCMI. (2014, October 13). http://dublincore.org/about-us/. Retrieved October 15, 2014, from http://dublincore.org/about-us/

Edmondson, R., & UNESCO. (2004). *Audiovisual archiving: philosophy and principles*. Paris: UNESCO.

First Archivists Circle. (2006). *Protocols for Native American archival materials.* [Salamanca, N.Y.]: First Archivists Circle.

Francis, K. D., & Liew, C. L. (2009). Digitised indigenous knowledge in cultural heritage organisations in Australia and New Zealand: An examination of policy and protocols.

Proceedings of the American Society for Information Science and Technology, *46*(1), 1–21. doi:10.1002/meet.2009.145046025

Gasiorowski-Denis, E. (2012, March 22). Adobe Extensible Metadata Platform (XMP) becomes an ISO standard. ISO. Retrieved from http://www.iso.org/iso/home/news_index/news_archive/news.htm?refid=Ref1525

Holton, G., Berez, A., & Williams, S. (2007). Building the Dena'ina Language Archive. In L. E. Dyson, M. Hendriks, & S. Grant (Eds.), *Information Technology and Indigenous People: Issues and Perspectives*. IGI Global.

Innes, P. (2010). Ethical problems in archival research: Beyond accessibility. *Language & Communication Language & Communication*, *30*(3), 198–203.

Internet Archive. (2014). FAQ. Retrieved from https://archive.org/about/faqs.php

Johnson, H. (2004). Language documentation and archiving, or how to build a better corpus. *Language Documentation and Description*, *2*, 140–153.

O'Neill, D. (2013, December 17). ID3 Introduction. Retrieved from http://id3.org/Introduction

Plichta, B., & Kornbluh, M. (2002). Digitizing speech recordings for archival purposes. *Michigan: Matrix, The Center for Humane Arts, Letters, and Social Sciences Online*, *7*. Retrieved from http://www.historicalvoices.org/flint/extras/Audio-digitization.pdf

TFADI, Commission on Preservation and Access, & Research Libraries Group. (1996). *Preserving digital information report of the Task Force on Archiving of Digital Information*. Washington, D.C.: Commission on Preservation and Access. Retrieved from ftp://ftp.rlg.org/pub/archtf/final-report.pdf

Toner, P. (2003). History, memory and music: The repatriation of digital audio to Yolngu communities, or, memory as metadata. In L. Barwick, A. Marett, J. Simpson, & A. Ha (Eds.), *Researchers, Communities, Institutions, Sound Recordings*. Sydney: University of Sydney: Open Conference Systems, University of Sydney, Faculty of Arts. Retrieved from http://hdl.handle.net/2123/1518

United States Copyright Office. (2014, July 1). Circular 56 - Copyright Registration for Sound Recordings. Library of Congress. Retrieved from http://www.copyright.gov/circs/

W3C. (2008). Table: Schemas used in XMP. Retrieved from http://www.w3.org/2008/WebVideo/Annotations/drafts/ontology10/WD/XMP.html

Wolf (Jr.), Michael Running. (2008). Building an Online Native American Archive: Problems and Premise. In *Intersecting interests: tribal knowledge and community research communities 2008 : a compendium of presentation articles.* (pp. 135–148). Missoula, MT: University of Montana. Retrieved from http://iers.umt.edu/docs/intersectinginterestsdocs/Intersecting%20Interests%20Compendium%20Final.pdf#page=135

Xiph.Org. (2013). The Ogg Container Format. Retrieved from https://www.xiph.org/ogg/

Xiph.Org Foundation. (2014a). OGG mapping. Retrieved from https://xiph.org/flac/ogg_mapping.html

Xiph.Org Foundation. (2014b). What is FLAC? Retrieved from https://xiph.org/flac/index.html