LANGUAGE DOCUMENTATION & CONSERVATION

Vol. 8 (2014), pp. 345-360
http://nflrc.hawaii.edu/ldc
http://hdl.handle.net/10125/24612

Developing a Living Archive of Aboriginal Languages

Catherine Bow, Michael Christie, Brian Devlin
Charles Darwin University

The fluctuating fortunes of Northern Territory bilingual education programs in Australian languages and English have put at risk thousands of books developed for these programs in remote schools. In an effort to preserve such a rich cultural and linguistic heritage, the Living Archive of Aboriginal Languages project is establishing an open access, online repository comprising digital versions of these materials. Using web technologies to store and access the resources makes them accessible to the communities of origin, the wider academic community, and the general public. The process of creating, populating, and implementing such an archive has posed many interesting technical, cultural and linguistic challenges, some of which are explored in this paper.

1. INTRODUCTION. During the era of bilingual education in the Northern Territory (1973 – 2000s), many books were produced at school-based Literature Production Centres in more than 25 languages. These materials, which are both widely dispersed and endangered, contain interesting and significant stories in Indigenous Australian languages, many with beautiful illustrations. As a result of policy and other changes, many of the materials produced for these programs are no longer in use, and in many places have been lost, damaged or, occasionally, deliberately destroyed. The goal of the Living Archive of Aboriginal Languages project[1] is to create a digital repository of this endangered literature and, with permission from the language authorities (usually original authors and illustrators or their descendants), to make the materials available to community members, researchers, and other interested parties through a searchable, online repository. The aim is to create a living archive with strong connections to the communities of origin.

The process of creating the archive has involved identifying and sourcing the books, scanning and digitizing them, and storing them safely. Once permission was obtained, the digital copies and any other related materials were then uploaded to the online archive so people could access them readily. The creation, population, and implementation of such an archive has posed a number of interesting challenges, as the project team endeavored to follow best practices in language archiving and to create a functional and user-friendly interface, while being culturally sensitive and responsive to community wishes. This paper discusses some of these technical, cultural, and logistical challenges and outlines what solutions were identified to resolve each of these sometimes-conflicting goals.

[1] The Living Archive project is a partnership between Charles Darwin University (CDU), the Australian National University (ANU), and the Northern Territory Department of Education, funded through a Linkage Infrastructure, Equipment and Facilities (LIEF) grant (LE120100016) from the Australian Research Council (ARC). The archive is accessible online at www.cdu.edu.au/laal/.

E-ISSN 1934-5275

2. BACKGROUND. In late 1972, the Commonwealth Government of Australia announced a policy of bilingual education for Aboriginal children in communities where traditional languages were spoken (Devlin 2009). Over the following decades, bilingual programs were gradually established in 20 remote communities in the Northern Territory, many with their own Literature Production Centre (LPC) located within the school. These LPCs were charged with the task of producing literature in key languages, and thereby providing the books and other resources needed to support vernacular and English literacy in bilingual programs. Some smaller schools had Literacy Centres which performed a similar function but without the printing equipment of LPCs. The aim of producing vernacular materials was to "flood the place with literature ... (for the) rapid and effective attainment of literacy in the vernacular" (O'Grady & Hale 1974:3), with the additional understanding that biliteracy programs assisted the transfer of skills to English literacy (Devlin 2011). Many senior people in communities supported bilingual education because it would allow their children to learn both traditional Indigenous knowledge and mainstream Australian knowledge, and (though often illiterate) they were committed to the possibility of preserving knowledge using whatever tools were available.

The selection of schools for bilingual programs was based on the existence of a dominant community language that had an established orthography, available specialist staff such as a linguist, and the interest and willingness of both Aboriginal and non-Aboriginal people to be involved in the programs. Of approximately 150 languages of the Northern Territory (Australian Bureau of Statistics 2011), around 25 were included in bilingual programs. Figure 1 shows the location of the LPCs and the languages included, while Table 1 lists the languages relating to each location.

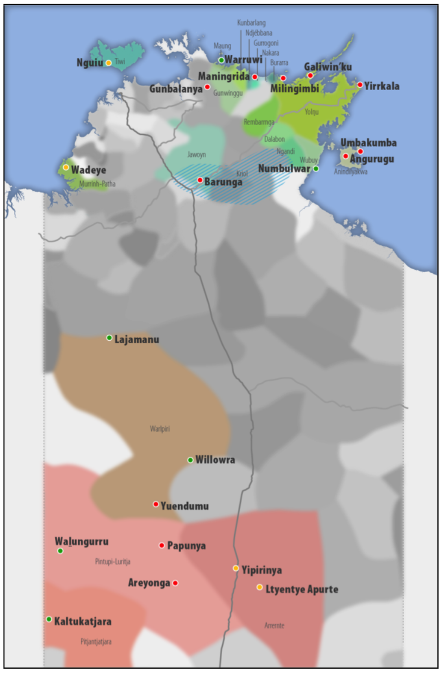

FIGURE 1. Map of languages and locations of bilingual education programs in the Northern Territory.

LOCATION(S) OF LPC/LC	LANGUAGE(S) OF INSTRUCTION IN THE SCHOOL	OTHER LANGUAGES IN COLLECTION	ESTIMATED NUMBER OF BOOKS
GOVERNMENT			
Angurugu, Umbakumba	Anindilyakwa		120
Areyonga, Kaltukatjara (Docker River)	Pitjantjatjara		220
Barunga	Kriol	Jawoyn, Dalabon (Ngalkbun), Rembarrnga, Mayali, Wardaman, Gunwinggu	350
Galiwin'ku	Djambarrpuyŋu	Gupapuyŋu, Galpu, Golumala, Ŋaymil, Dhaŋu, Rirratjiŋu, Warramiri, Wangurri	330
Gunbalanya (Oenpelli)	Kunwinjku		50
Maningrida	Burarra, Ndjébbana (Gunibidji)	Gurrogoni, Kuninjku, Nakara, Wurlaki, Djinang, Dalabon (Dangbon), Kunbarlang	400
Milingimbi	Gupapuyŋu	Djambarrpuyŋu, Liyagalawumirr, Gumatj, Yan-nhaŋu, Maḏarrpa, Dhalwaŋu, Rirratjiŋu, Liyagawumirr	250
Numbulwar	Wubuy (Nunggubuyu)	Ngandi, Mara, Ritharngu, Wagilak	160
Papunya, Waḻungurru (Kintore)	Pintupi Luritja		300
Warruwi	Maung		350
Yirrkala	Gumatj, Dhuwaya	Dhuwala, Dhalwaŋu, Djapu, Maḏarrpa, Rirratjiŋu, Wangurri	320
Yuendumu, Lajamanu, Willowra	Warlpiri		500
CATHOLIC			
Santa Teresa (Ltyentye Apurte)	Arrernte (Eastern)		150
Nguiu (Wurrumiyanga)	Tiwi		300
Wadeye	Murrinh-Patha		250
INDEPENDENT			
Yipirinya	Eastern Arrernte, Western Arrernte, Warlpiri, Luritja		250

TABLE 1. List of locations and languages of bilingual education programs in the Northern Territory.[2]

Subsequent policy changes led to the demise of bilingual education in most of these communities (Devlin 2009; Simpson, Caffery & McConvell 2009), though a few schools have retained some language and literacy programs. This has meant the closure of many LPCs, while a few remain active and others have been repurposed (usually as photocopy rooms or storage areas).

The impetus behind the Living Archive project is the resulting endangerment of the materials created in the LPCs, due to fluctuations in policy, and sometimes school leadership. In some communities, all the copies of particular books have disappeared or been destroyed. In at least two locations there are tales of principals clearing out whole collections of books and having them burnt or taken to the tip (landfill). Extant books which are available in various libraries (such as Australian Institute of Aboriginal and Torres Strait

[2] The second column of this table lists the primary language(s) of the bilingual programs, while the third column includes both language group names (such as Dhuwala) and individual clan languages (such as Djapu) in which additional materials were produced.

Islander Studies (AIATSIS), the National Library in Canberra, Northern Territory Library, and academic collections such as the Australian National University, Charles Darwin University, and Batchelor Institute libraries) are often stored in special collections of rare books available only for registered users and not for loan, and are thus largely inaccessible to the public. While some materials have been digitized, those digital versions were not held in any centralized location, making discoverability and access quite difficult.

In addition, the materials have in many cases become isolated from their communities of origin and particularly from the various language groups where they were produced. Another goal of this project is to establish unique social and technical possibilities for increasing the participation of remote Aboriginal knowledge authorities in areas of significant research by connecting (or re-connecting) the materials with the owners and with researchers interested in collaborating with relevant people in communities. The project also aims to help Aboriginal people gain access to a part of their intellectual history, to allow them to mobilize epistemologies, histories, environmental knowledge, and narrative practices towards cultural and linguistic sustainability for themselves and others. This trans-disciplinary project crosses into a number of different fields, including language documentation and description, language endangerment, digital archiving, and indigenous epistemologies and pedagogies. While some of these fields have developed 'best practice' codes, others are still emergent; and in some cases there may be inherent contradictions between the requirements of each field.

3. MATERIALS. A wide range of materials was developed in these LPCs for use in the classroom and for the wider community. They included teaching materials, literacy primers, children's stories, stories of local cultural significance such as environmental knowledge, traditional practices, oral literature, ethno-botany, history, non-sacred versions of Dreaming or Creation stories, experience stories, instructional manuals, cautionary tales, and many others. Many books were based on recordings of old people telling stories; others were developed by students in the classrooms. The work of converting an oral narrative to a written one is an intricate process involving complex and subtle decision-making on both linguistic and literary grounds. Literature production workers, usually Indigenous native speakers literate in their own languages, engaged with teacher-linguists and literature production supervisors to create and publish these resources.

The books range in size, length, content, and print-runs, making it difficult to describe a 'typical' item. While the majority range between A5 and A4 in size,[3] there are 'Big Book' formats up to A2 size, and small readers in A6 size. In general the books contain between 10 and 30 pages, though there are examples of longer books. Some books have one word per page; others have long texts of several thousand words. The majority of the books contain illustrations, usually by local artists, and range from simple line drawings to detailed hand-painted images. Literature Production Centres were equipped with offset printers, and later the introduction of color printers and desktop publishing programs allowed for more sophisticated production techniques. Around 100 copies of each item were generally published, usually with light card covers, folded and stapled.

Often materials published in one center were translated and adapted into other languages for different communities. Some stories were translated from familiar children's stories in English, sometimes adapted to local situations. For example, the well-known children's story *Are you my mother?* by P.D. Eastman was originally published in the US

[3] http://www.papersizes.org/a-paper-sizes.htm

in 1960, about a hatchling bird asking other animals (cat, chicken, dog, cow) in turn if each one is its mother. A local adaptation in Ndjébbana language replaces the bird with a wallaby and the other animals with a buffalo, dingo, crocodile, emu, etc. Other local versions of this story exist in several languages in the collection, as well as other popular, non-local stories such as *Three billy goats gruff*. Other stories were translated with no adaptation to local contexts, such as *Phantom* comics in the Maung language of Goulburn Island. Some collections include a range of 'paste-over' books, where copies of books in English were translated into the local language, but rather than being republished in the language, the translated sentences were simply glued over the English text on the page. Similarly, a series of blank readers by Science Research Associates with pictures and no text was distributed to schools with bilingual programs in the early 1980s, and local language versions of these books were produced. While the content of these stories may have limited relevance to local contexts (e.g. stories of pandas and elephants), the language content is of some interest.

As noted above, the majority of books in the collection are illustrated, some with simple line drawings, others with intricate paintings by local artists. Some illustrations were digitally enhanced with the introduction of desktop publishing software in the LPCs in the late 1980s. In some cases, pictures drawn by local schoolchildren were used, while others include photos taken around the community. These were often used in the genre of experience narratives, for example *Al-mangiyi Garri-meyh* ('We got a longneck turtle') in the Jawoyn language, and instructional books such as *Ngirramini ngini pamijini* ('How to make an armband') in the Tiwi language. Identifying genres for these resources for categorization purposes is a challenge. What indigenous people may identify as 'history' could be considered 'folktale' by non-indigenous readers. Categorizing Aboriginal literature into European literary genres may in fact cause more problems than it solves, because it undermines traditional classifications of modalities. Photo books may also be problematic where communities do not allow photos of people who have passed away, so this requires careful negotiation with key authorities before such material is made public. While there are hundreds of such items in a collection, each story is the product of a unique set of circumstances. Limits of time and resources make it difficult to address each individual item in the collection with the appropriate authorities, and so it is the easy cases that make their way to the front of the line. However, mis-steps in this space can have consequences for the entire project, risking rejection and censure from the communities whose materials are being archived (see McConvell 2000 for further discussion on traditional restrictions in modern contexts).

The material in the Living Archive is largely limited to printed matter. However, where audio or video materials are available digitally and correspond to books in the archive, these can be linked to book records available on the website. This limitation is largely due to resource capacity, rather than a judgment about the value of such materials in language documentation and preservation. The Living Archive is designed to be extensible, with the desire that future funding arrangements may allow the inclusion of valuable and important materials in these formats, as well as new materials being produced in the few still-functioning LPCs.

4. THREE-PART PROCESS. The process of compiling the archive involved three main activities, which sometimes occurred concurrently.

4.1 IDENTIFY. Not all of the books produced by, or held in, the LPCs had been catalogued, so one of the first tasks for the project team was to create an inventory of materials to be collected for digitization. All books published in Australia are normally submitted under a legal deposit requirement to the National Library and state libraries, which catalogue them using appropriate metadata relating to each item. These items are easily accessed through the National Library of Australia's Trove website[4], and for this project a download of metadata for around 3000 records was procured from there early in the inventory process. Some information in the Trove catalogue, however, is incomplete or erroneous, which is hardly surprising when several items have no accompanying English gloss or metadata, so guesses had to be made by library cataloguers as to the identity of the author or illustrator or even language. In some cases the language of a book is simply listed generically on Trove as 'Australian language' or erroneously recorded as 'English.'

However, not all titles were submitted to these official sources, and for various reasons it was not always considered important to include the names of authors or illustrators, or dates of publication, etc. Some literature production supervisors and teacher-linguists were meticulous about recording all these details, providing a rich source of information about each title. In other cases, the title was the only metadata available for a book itself (and some items did not even have clearly identifiable titles, or a different title on the cover page from that on the title page). Some metadata in the Living Archive was added from external sources; for example, an illustrated book translated from another language that does not include the name of the illustrator. Local knowledge was used, where people in community or school staff recalled or recognized material produced by specific individuals. Information from secondary sources such as these is indicated by the use of square brackets in the metadata, to conform to best practice in digital archiving.

While some communities were diligent about the storage and archiving of materials, in other cases, little or no archiving had ever been done. Some materials were carefully stored in labeled boxes, or available to browse on display stands in a library or resource center; in some places they were held in a compactus in a storeroom, or unceremoniously dumped in a dusty shed or storage area. In at least one site, all the materials had been destroyed, except for a small collection salvaged by the local missionary.

Information from the National Library catalogue, plus spreadsheets of metadata from LPC records compiled by a Masters student at Charles Darwin University in 2003, were used as a starting point to source materials from LPCs, libraries, and private collections. The project manager for the Living Archive project visited many of the schools that had bilingual programs to update the spreadsheets and to find clean copies of the materials for digitization. Other goals of the visits were to discuss and promote the project with key people in the schools and in the community, and to seek the approval of authors, illustrators and others involved in the creation of the materials to upload them to an open website.

The metadata available included information about titles (including translated titles, subtitles, alternative titles), creators (such as author, illustrator, translator, editor, photographer or any other contributor), and publication details (date, place, ISBN if available). Additional information was also included in the spreadsheets but not on the final archive, such as where the item was located (e.g., in a library, LPC, or private collection), and its status in the archive (e.g., if a digital version was available, if a text file had been created, etc.). Where items were listed on the Trove catalogue, a reference number was also included, in order to facilitate sharing data with the National Library in the future. Decisions

[4] www.trove.nla.gov.au

were made as to what metadata was needed to keep track of the objects (and which would not appear in the final archive), what metadata would be most useful for language owners (and which should therefore be prioritized), and which might be of interest to archivists and should be preserved (though hidden in the first appearance of records). These decisions were sometimes reconsidered as the project developed, leaving some inconsistencies and gaps to be rectified later.

There were several challenges associated with creating the inventory of materials. As noted earlier, many items contained minimal metadata, or sometimes contradictory information. Naming practices for people are a common concern to librarians, particularly with cross-linguistic variants following different rules, and there does not appear to be a standard convention for handling Indigenous Australian names. Western and Aboriginal names are sometimes variably combined, and sets of subsection or ’skin names’ may also be employed on occasion. For example one contributor may be listed variously as Mary Nakamarra Smith, Mary Smith, Mary Nakamarra, Nakamarra Smith or Mary Smith Nakamarra.

Cultural practice regarding the naming of deceased people has not been a significant issue in recording metadata; however, care was taken when talking with Indigenous people about the deceased. Special characters also introduce some variability in naming and this affects the standardization of names; for example, the Yolŋu family name Munuŋgurr is sometimes listed as Mununggurr (or, less accurately, as Munuŋurr or Munungurr). Metadata relating to creators of the resources sometimes include a nominal suffix meaning ‘made by’ or ‘from’ (such as *-wuŋu* and its variants in Yolŋu languages, or *-rlu* in desert languages), which needs to be stripped from the name in the metadata record. Besides the practical challenges variant names cause for the archiving project, there are larger issues of mismatch between western and Indigenous naming practices, regarding reinforcement of patriarchal connections, expectations of consistency in names as opposed to people changing names, and having several names which go in and out of fashion. Best practice in cataloguing involves faithfully recording metadata as it is listed in the original item, but accurate search functionality requires consistency. This issue is not unique to the Living Archive project, and affects all librarians and cataloguers. The solution devised for this project was to list a contributor’s name exactly as it is displayed in a book, and to link this name to a stable four-digit number, which would connect an individual author to any variation of their name (including erroneous spellings). This situation relies on local knowledge sources to inform this process, with the hope that errors can be identified and edited.

This wide range of materials created a dilemma for the archive developers in deciding what materials to include or exclude. The original focus of the project was preservation of original materials written and illustrated by local Indigenous authors and artists, which would be of anthropological, cultural, and linguistic interest. However, the wider field of language documentation maintains that any and all material produced in these endangered languages have inherent value. Translations from English have some cultural value, for example, because they demonstrate how native speakers of Aboriginal languages translate various texts and reconstitute them in written form. The collections also include teaching materials, such as primers, word lists, puzzles, coloring books, teachers’ resources, and occasionally linguistic notes, which could be included from an educational perspective, both as historical records and materials that can be reused or reconstituted for teaching programs. Health manuals, reports of local events, local newspapers, etc., also form part of the potential corpus, and are included if time and resources allow.

While the focus is on materials published by the LPCs themselves, additional materials are available for inclusion in the digital collection, such as those produced by the Summer Institute of Linguistics (SIL; now AuSIL – the Australian Society for Indigenous Languages), materials developed by students at Aboriginal Languages Fortnight for the now-defunct School of Australian Linguistics, and occasionally commercial publications. Discussions about copyright often impact the decisions as to which materials can and should be included, so that a number of items have been digitized for preservation, even if they are not made public on the website.

Language names are another area of complexity. While the International Organization for Standardization (ISO) gives standard three-letter codes to represent language names, ISO 639-3[5], it lacks codes for some languages in the Living Archive collection. In some cases this relates to Indigenous connections to a language being rooted in clan affiliations, for example what ISO 639-3 lists as Dhaŋu language (dhg) encompasses the clan languages Galpu, Golumala, Ŋaymil and Warramiri. Books published in these languages are identified in the metadata as being in Galpu or Golumala, for example, but to conform to ISO standards they need to be listed as Dhaŋu. This results in different translations of a book appearing to be in the 'same' language, and creates complexity in terms of searchability. Alternate names and spelling of language names (for example, the fact that Djeebbana (djj) is the ISO 639-3 name for the language labeled as Ndjébbana (occasionally also as Gunibidji) in all the materials created at Maningrida LPC) adds further layers of difficulty. Best practice for linguists (Simons & Aristar Dry, 2006) may differ from best practice for librarians (using metadata as it is printed in the resource itself) and best practice for the Aboriginal owners of the languages, which is still emergent. It is possible to submit change requests for ISO codes to better reflect the complexities identified on the ground, a task that the Living Archive project team has initiated.[6] Even identifying the language(s) of a particular item can be challenging; for example, some important stories of ancestral work very deliberately and cleverly present particular totems and ancestral tricksters speaking a series of different dialects to make points about ownership, rights, position, etc. These complexities, which are highly significant in an Aboriginal polity, are not easily accommodated in a conventional archive by a single language label.

Geographic naming practices also affected communities identified in the archive. Some LPCs have changed names over time (e.g. Santa Teresa Literacy Centre was later renamed Ltyentye Apurte Literature Production Centre; another centre is listed variously as Literature Production Centre Elcho Island, Galiwin'ku Adult Education Centre, Galiwin'ku Literature Production Centre and Shepherdson CEC Literature Production Centre); some places changed name (Oenpelli became Gunbalanya, Bamyili became Barunga, the place currently called Wurrumiyanga is listed in the archive as Nguiu), or are known by two names (such as Docker River and Kaltukatjara). A synonym list was devised to allow for various user inputs to produce correct results to search queries in the Living Archive.

The tension in recording 'accurate' metadata is also apparent in the use of International Standard Book Numbers (ISBNs), which should be unique identifiers of distinct items. However, in this archive there are examples of the ISBN listed in the book itself actually referring to the original item from which this was translated, for example *Djalwarra Ŋarraku Yarrar'yurr*, published by the Galiwin'ku Literature Production Centre, lists its ISBN as that of the original publication, *The Day I Split My Pants*, published by Nelson. In other cases, revised editions of an item or even translations into different languages share the

[5] http://www-01.sil.org/iso639-3/codes.asp

[6] http://www-01.sil.org/iso639-3/chg_requests.asp

same ISBN. The practice of the Living Archive project is to include ISBN numbers where available, but avoid where possible the confusing duplication or inappropriate assignment of these identifiers.

The archive is grouped into 16 ‘collections,’ which may be based on a location (associated with a number of languages) or a language (with a number of locations). For example, the collection from the community at Maningrida includes materials in a number of languages of that region (Ndjébbana, Burarra, etc.), but does not include materials that were published there if they originated in a different community. Goulburn Island had a Literacy Centre but no printing facilities, so the Maung language materials developed there, but printed at Maningrida, are included in the Maung rather than the Maningrida collection. On the other hand, the Warlpiri collection includes materials published at the Bilingual Resources Development Unit (BRDU) in Yuendumu, as well as those developed in the bilingual programs at Willowra and Lajamanu schools. The BRDU also published some materials in Pitjantjatjara and Pintupi-Luritja, which are included in the respective collections of those languages. While there were some areas of overlap or uncertainty (for example there are Rembarrnga language materials published in both Maningrida and Barunga LPCs), the 16 collections proved a relatively straightforward and helpful way to manage the archive, while being largely opaque to the end user.

Corporate and local memory is an invaluable tool that was tapped to understand the history of the literature in each community, and the various decisions made about its use and storage. The project exposed and in some cases unraveled many interesting issues from the corporate memory of both Indigenous and non-Indigenous staff and community members regarding the history of literature development and the storage and usage of the materials over several decades. Such valuable information was used to update and supplement the metadata spreadsheets and to inform decisions about the archive. Some strategic connections unlocked other small, largely forgotten collections. For example at Umbakumba on Groote Eylandt, most people reported that no materials had been saved anywhere in the community. However, a chance conversation led to a former teacher who had stored a small collection of paste-over books in a storage room in the school, which had remained hidden for many years, and was delighted that someone was finally doing something with these items.

4.2 DIGITIZE. Materials collected for digitization were handled in a variety of ways within the project. The library at Charles Darwin University has a digitization kit with two digital SLR cameras mounted on either side of a V-shaped book cradle with its own lighting, linked to a computer with image processing software. The National Centre of Biography at the Australian National University used a similar setup to digitize a number of items sent from Central Australia, while the Northern Territory Library also allowed the project team to use their mobile digitization kit. In some communities where scanning could be done on site, staff used the school’s multi-function printer; however, this did not produce optimal results. Flatbed desktop scanners were used in the Living Archive office and also provided to some communities for in-house scanning. When scanning was done in remote schools, the image files were shared using Dropbox[7] to be processed in Darwin. Adobe Lightroom imaging software[8] was used to crop and edit the images, and the raw images were saved as uncompressed Tagged Image File Format (TIFF) files with a minimum resolution of 400 pixels per inch (ppi).

[7] https://www.dropbox.com/

[8] http://www.adobe.com/products/photoshop-lightroom.html

The archive also inherited some legacy materials from digitization projects that predated the current project. In Barunga, for example, an earlier project established by a principal (with support and funding from the Northern Territory Department of Education and Katherine Group Schools) had digitized all material held in the LPC prior to its closure. This collection was stored on a 4-terabyte hard drive, with each individual page carefully labeled with dozens of separate spreadsheets containing detailed metadata about each image. This compilation included drafts, masters, notes, and materials from other collections produced in other communities. The Living Archive project was given access to these digital materials, and selected appropriate items from which to collate and create PDFs (Portable Document Format), then followed up with the community to seek permission to include those materials online at the Living Archive website.

Given the variety of sources from which digitized materials were obtained, it was not always possible to maintain quality control across the archive. Materials not digitized within the Living Archive project were of varying quality, and where possible, some images were edited (e.g., cropped, straightened, lightened); however, some items saved directly as PDFs from multi-function printers could not easily be improved. In a few cases it was possible to rescan items of very low quality (if the hard copy was available); however, due to limits of time and resources, it was sometimes necessary to accept a poor quality digital version than have none at all. Higher quality versions can be uploaded if they become available later.

In some locations, access to hard drives or servers revealed other digital versions of materials for the archives. From Yuendumu and Maningrida for example, complete final PDF versions of the latest editions of many items were made available to the project, which avoided the need to scan the corresponding printed books. From Nguiu and Numbulwar, earlier incomplete versions of some items were available, and some could be reconstituted using desktop publishing software tools. Some items were available in different formats (such as ‘big book’ editions for group reading) or revised versions (e.g., with digital coloring of original artwork). Some variants had only a few spelling or punctuation differences, or a different-colored cover page; others had undergone more significant changes. Best practice requires separate entries for discrete items, which required careful checking to identify these differences.

It was often necessary to distinguish between books with the same title – which may be different books on the same topic, or the ‘same’ book in different formats. For example, there are three books in the Arrernte collection entitled *Aherre* (‘Kangaroo’) with no additional metadata for author, illustrator, year, etc. Two books have identical pictures on the cover, and on close inspection were identified as different versions (one a ‘big book’ version) of the same book, following the pattern *Kangaroo, kangaroo what do you see?* The pictures and text in these two are almost but not quite identical, and there is no distinguishing metadata such as year of publication. The Related Items function in the interface allows such associated resources to be linked. There are other cases that involve digitized versions of exactly the same book coming from different sources. Each case requires careful viewing to determine their similarity or difference with respect to other items. A less common situation involves changes in orthography. Two versions of an Anindilyakwa book tell the story of a dog and a pheasant – the original version from the 1970s uses an older orthography (with underscored letters), while a colored, handwritten version from the 1990s is in the current orthography. The titles are different, and there is limited metadata, so two separate records are created and linked using Related Items.

The task of creating text files for each of the items in the collection required the use of Optical Character Recognition (OCR) software, and Abbyy Fine Reader[9] was selected for this task. The process was often very problematic, due to the variable quality of the original materials, including those that were handwritten, faded, of poor quality printing, or with text overlapping with images. The software also had difficulty handling special characters used in Australian Indigenous languages. Additional Unicode characters (such as d̲, ŋ, ä, r̲, é, etc.) were added to the software's internal dictionary. However, these often needed to be retyped, as they were not always recognized in the original texts. Even in some cases where the font appears clear and consistent (e.g., in the Tiwi language materials), serifs, italics and other features affected the interpretation of the letters and required much revision to the original processing. In some cases it was simpler to type the complete text from scratch rather than OCRing. On occasion, the OCR software distorted the pages being processed – turning straight edges of images wavy, or creating different page sizes within a single book. An additional problem was that once a file was OCRed and saved as PDF, no further edits could be made; any further corrections would necessitate re-processing and re-editing the entire document from scratch. Unfortunately, this problem was not identified until after hundreds of titles had been OCRed and saved as PDFs prior to final quality control checks, including some saved with no corrections to the initial OCR reading. As a result, it was decided that the accompanying text file should be as accurate as possible, even if the presentation PDF version of the document contained OCR errors embedded in the hidden text. In some cases this meant resaving the PDF as an image only, rather than as a word-searchable PDF. This allowed the project to retain the searchability of texts within the collection, without spending additional time reprocessing hundreds of items that looked fine but were not searchable. When OCR software was used, careful checking was required for each line of text, which was challenging for those not familiar with the language. It is hoped that the expertise of local people who are literate in these languages can be drawn on to check and edit these texts. This opens up possibilities for the Living Archive in the future, to try crowdsourcing as a means of correcting items found to contain errors.

The OCR process raised other issues; for example, in some cases there were typographical or spelling errors in the original text. This draws on a larger question of whether the integrity of the original publication should be valued over the accuracy of the language data (a question which also relates to metadata). On the one hand, since the texts will be used for searching, lexicography, and linguistic analysis, then preserving errors in them anywhere other than the scans serves no purpose. On the other hand, best practice for archiving says that the authenticity of the original artifact should be maintained, and any alterations can be negotiated separately. Some books digitized for the project contain handwritten annotations, often corrections or notes on different spellings, but there is no indication of the source or validity of the annotations, so it is unclear if they should be included in the text file or not. One solution adopted in this project was to attach additional text files to each record, an 'original' file with the text included as written, complete with any erroneous spelling or other non-standard presentation, plus a 'corrected' file with any corrections and attribution of who made these corrections, if such information is available.

To conform to standard archiving practice, both preservation and presentation formats were created for each resource. Preservation formats allow high quality representation of the digital objects, but the very large file sizes make them unsuitable for web delivery. Tagged Image File Format (TIFF) was chosen as the format for preservation copies

[9] http://www.abbyy.com.au/finereader

of the image files, while the presentation versions use PDF, as a widely used, flexible and (ideally) sustainable means of delivering both text and images in a commonly-used and well-supported format. In addition, the text was saved in plain text (Unicode-compliant), and the covers saved as smaller Joint Photographic Expert Group (JPEG) images (800dpi) for quick access via the website. The preservation versions are stored on the Charles Darwin University digital repository with the metadata and presentation versions; however, they are not directly accessible through the online interface, but rather available through an application process. There is ongoing discussion within the project team about mirror or backup versions stored in other locations, and how these may be accessed and maintained.

The result of this digitization process is that each item in the archive has at least three formats: PDF, TXT (Unicode) and JPG (cover page only), and most also have TIFF (as preservation files of larger size). Efforts were made to limit the PDF size, to both facilitate quick downloading and to assist users with limited download capacity. Where other formats exist, such as audio or video recordings of the stories, or 'talking books' combining audio and video, these are linked as related items to the appropriate record and can be retrieved through the interface. While creating additional formats is not among the goals of the current project, any such items that are shared with the archive (such as e-books) will be included.

4.3 PUBLISH. The third step in the process is to upload the digitized materials to the online repository. The Living Archive is hosted on the Charles Darwin University Library's institutional digital repository for research and teaching materials (known as eSpace). This repository runs on Fez[10] as a web front-end for Fedora Commons open source repository software.[11] Working closely with the Digital Collections coordinator and technical support staff, a Document Type Declaration (DOCTYPE) was developed to map Metadata Object Description Schema (MODS, a bibliographic description schema based on XML and used widely in libraries) fields to appropriate categories as identified by the Open Language Archives Community (OLAC).[12] The use of OLAC metadata standards accords with the project's goal of being compliant with the current best-practice standards for language resource descriptions in ways which support preservation and promote discoverability of the resources in the archive (Bird & Simons 2003).

5. PUBLIC INTERFACE. The other aspect of the publication process is ensuring public access to the archive. As the Fez/Fedora online repository is accessible only to staff and students of Charles Darwin University, a separate interface was required to overlay on to the Fez/Fedora repository. A contractor was engaged to develop this, with instructions to build a site that could be accessed on multiple devices and operating systems. The site had to be accessible not just to the academic community or highly competent computer users, but also to those in remote communities represented in the archive, especially people who may have limited access to and familiarity with technology, limited connectivity, and often lower text-literacy levels than other users. To give preference to such users, it was decided to build a highly visual interface, with access to languages and locations via a map, and to display cover images of the materials for selection, instead of restricting search and display results to text. It is thought that privileging Indigenous users of the archive would not

[10] http://fez.library.uq.edu.au/

[11] http://www.fedora-commons.org/

[12] http://www.language-archives.org/

exclude others, but would provide an alternative means of accessing the huge amount of information included in the archive, without requiring either high text-literacy or high technical skills. While access to technology in remote areas of northern Australia is improving, and in some cases is excellent, with both children and adults regularly accessing online services and using smartphones and tablets for various purposes, it could not be assumed that all users would have the same level of access. Use of older browsers or slower machines requires options for simpler and faster access to the materials, without compromising the overall quality of the site.

The site[13] incorporates a map interface that allows users to click directly on either a language region (marked in color) or a specific location (indicated by an inverted tear drop) of the LPC (see figure 1). More traditional Search and Browse functions are also available, displaying cover images of items returned by such queries. Additional clicks will display entire PDF files and full text files embedded in the page and also available for download. Where audio or other related items are available, these are visible on the record view, with simple swiping or scrolling between records and user-controlled options for sorting and on-screen display. The architecture is extensible so that additional languages, locations and metadata fields can be included as necessary.

6. PERMISSIONS. Copyright of most of the materials in the Living Archive collection belongs to the Northern Territory Department of Education which, as a partner in this project, has approved the uploading of all these materials to the project's public website. Beyond this, however, it was agreed that the original Indigenous creators of the materials should be included in the negotiation, to be informed about the archive, and be given the opportunity to decide whether or not their items can be made public online. Since the materials were developed for use in the school community, there are no examples of secret or sacred content. There may, however, be preferences for some material to be kept private. The project does not wish to assume compliance, but to work with the community members to decide which materials should (or should not) be included in the archive. In many cases, the people involved had passed away, so efforts were made to identify an appropriate family member who could give permission on their behalf.

To facilitate this process, permission slips were created with a brief explanation of the project, and the option for people to sign for "all materials I was involved with creating" and to identify any materials they did not want included. An updated version of the form included room to nominate family members being signed for, as well as any other names people may be, or have been, known by. The project manager, members of the project team, and Department of Education staff took these forms on site visits and met with as many people as possible to discuss the project and gauge their approval. Where possible, uploaded materials were shown on beta versions of the online repository, to demonstrate how the materials would appear to the public. It was explained that "no one will be allowed to sell or buy the materials," however it was acknowledged that it would be more difficult to challenge any copyright breaches outside of Australia, as internet behavior is more difficult to regulate outside the country (a click-through warning is also included on the site which acknowledges these terms and conditions). Any issues that arose were addressed, and it was emphasized that there was no obligation to give permission. It was also explained that the benefit of having the materials available to a wide audience to appreciate and enjoy should outweigh any risks of exploitation and abuse. To date, no one has refused permission

[13] www.cdu.edu.au/laal

for their own or their family's materials to be included in the archive. As permission signatures were collected, more materials were shifted to the public collection visible on the website.

Of particular concern were materials containing photos, due to cultural sensitivities about showing images of people who have passed away. Photos of children should also be handled sensitively, and many books include images of school children involved in various activities. Another challenge was handling materials with multiple authorship, such as books produced by a whole class, where it was virtually impossible to track down each member of the class (or their family members) to approve uploading. Community consultation was required to decide for each community and (where possible) each book. Another concern was titles with limited or no metadata – where there is no mention of who wrote or illustrated an item, this does not mean it can be uploaded with impunity, but should be withheld until appropriate creators are identified and contacted. Copyright law considers these 'orphan items' which require a risk management approach, involving attempts and invitations to determine authorship and a 'good faith' notice inviting interested parties to come forward should they have concerns about what is being done (Australian Copyright Council 2012). There were also situations where a book from another language (e.g., popular children's books in English, as exemplified earlier) was translated or adapted into the local language, with or without attribution to the original creators. In some cases the Department of Education sought and received permission to use external materials (such as the Science Research Associates materials) but in others, because materials were only for in-house use, such permissions were not as carefully negotiated. The project team is currently in discussion with the Department of Education regarding allowing a Creative Commons license to be used on the website. Similar negotiations are in progress with non-government school authorities such as the Catholic Education Office.

7. CONNECTION WITH COMMUNITIES. The goal of this project was to develop a 'Living' Archive, which would be more than simply a repository of language materials, but with vibrant, continuing connections to the places and people of origin. To avoid perpetuating the notion of a group of non-Indigenous academics taking language and cultural materials and locking them up in a database that is inaccessible to people in remote communities, the project team's intention was to include people that the materials were originally developed by and for in the process of archiving. While the technical aspects were largely handled by the project team, the socio-cultural aspects involved much discussion and negotiation with people in these communities. Seeking their permission and explaining the project during site visits was one crucial aspect, allowing local Indigenous people to make decisions about what should or should not be included in their collections. In communities where large numbers of items were identified for archiving, desktop scanners were provided and local Indigenous people employed on a casual basis to continue scanning items that were not taken to Darwin. In Galiwin'ku for example, a local Yolŋu person was engaged to track down original creators of materials or their descendants to seek permission to upload materials into the archive. It is hoped that future work on the archive will enable additional means of engaging communities in enhancing and customizing the collections.

Another way of including local people in the development of the archive is in the correction and enrichment of materials. Since the process of OCR is not error-free, and the casual staff at Charles Darwin University working on this process were not literate in any of the languages of the archive, literate speakers of these languages can be invited to read through the texts and edit or correct any errors. Enrichment can also come in the form of audio-recording speakers reading aloud from the books, or recording additional stories about the books. For example, a senior authority in Pintupi-Luritja language was recorded reading some stories from the archive, then discussing how the language had changed since they were written. The creation of talking books is an activity that school children could engage in, to record and animate stories included in the archive – the technology is simple to use and readily available in some communities, and would engage local children in developing dynamic versions of otherwise static materials. Holton (2012) discusses some of the unforeseen potential uses of materials stored in language archives.

8. CONCLUSION. The Living Archive of Aboriginal Languages has been developed to preserve and allow access to rich resources of significant linguistic and cultural importance. The project has established the infrastructure on which to store such a wide variety of materials, and populated it with hundreds (expected soon to be thousands) of resources, and made it publicly accessible on the Internet. The challenge now is to engage local language authorities in collaborative activities to expand, enhance and customize the collections, and to engage students and researchers from around the world with the archive contents and its owners. A language archive should not be a dry, academic, static repository, but has the potential to become a living, growing organism linking people, places, stories, languages, cultures, and epistemologies across time and place.

References

Australian Bureau of Statistics. 2011. 1267.0 Australian Standard Classification of Languages (ASCL). http://www.abs.gov.au/ausstats/abs@.nsf/cat/1267.0. (22 September, 2014.)

Australian Copyright Council. 2012. Orphan Works (Information Sheet G101v04). www.copyright.org.au/admin/cms-acc1/_images/1549612446523924db2ad24.pdf. (22 September, 2014.)

Bird, Steven, & Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language* 79(3). 557-582.

Devlin, Brian. 2011. The status and future of bilingual education for remote indigenous students in the Northern Territory. *Australian Review of Applied Linguistics*. 34(3). http://www.nla.gov.au/openpublish/index.php/aral/article/viewFile/2277/2738. (22 September, 2014.)

Devlin, Brian. 2009. Bilingual education in the Northern Territory and the continuing debate over its effectiveness and value. Paper presentation at AIATSIS Research Symposium: Bilingual Education in the Northern Territory: Principles, Policy and Practice 2009, Canberra, Australia. http://www.abc.net.au/4corners/special_eds/20090914/language/docs/Devlin_paper.pdf. (22 September, 2014.)

Holton, Gary. 2012. Language archives: They're not just for linguists any more. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, and Paul Trilsbeek (eds.), *Language Documentation & Conservation Special Publication No. 3: Potentials of Language Documentation: Methods, Analyses, and Utilization*, 105–110. http://hdl.handle.net/10125/4523.

McConvell, Patrick. 2000. Two-Way Research Resources for Indigenous Languages: Positioning Resources in the GARMA. *Papers from the Workshop on Web-Based Language Documentation and Description*. Philadelphia: Institute for Research in Cognitive Science (IRCS), University of Pennsylvania.

O'Grady, G. & K. Hale. 1974. *Recommendations concerning bilingual education in the Northern Territory*. Darwin: Department of Education.

Simons, Gary & Helen Aristar-Dry. 2006. Good, Better, and Best Practice. *Proceedings of the Deutsche Gesellschaft für Sprachwissenschaft*. Bielefeld. http://emeld.org/documents/Bielefeld-Dry-Simons.pdf. (22 September, 2014.)

Simpson, Jane, Jo Caffery, & Patrick McConvell. 2009. Gaps in Australia's Indigenous language policy: dismantling bilingual education in the Northern Territory. *AIATSIS Research Discussion Paper* (24). http://www.aiatsis.gov.au/_files/research/dp/DP24.pdf. (22 September, 2014.)

Catherine Bow
cathy.bow@cdu.edu.au

Michael Christie
michael.christie@cdu.edu.au

Brian Devlin
brian.devlin@cdu.edu.au