Supporting Australian Indigenous languages from textual sources[1]

Nick Thieberger
School of Languages and Linguistics, University of Melbourne, Australia
ORCID: 0000-0001-8797-1018

Sophie Lewincamp
School of Languages and Linguistics, University of Melbourne, Australia
ORCID: 0000-0003-1783-627X

Marco La Rosa
School of Languages and Linguistics, University of Melbourne, Australia
ORCID: 0000-0001-5383-6993

***Abstract*— For many Indigenous languages there are few records, and the earliest written sources are witnesses of the language as spoken before major changes that were typically brought about by colonial expansion. In Australia, as a result of settler aggression, removal of Indigenous children, and displacement of the original inhabitants from their land, many Indigenous languages are no longer spoken. In this situation, the earliest sources that record aspects of these languages become all the more valuable as resources for relearning the languages. However, as paper documents in a single repository, they can be difficult to access and to use. We report on our project to take such manuscripts, convert them to text, and to create a platform in which they can be found and used. We allow for various input formats for the text and store it and images as Research Object Crates (RO-Crates). We use Amazon Textract for Optical Character Recognition (OCR) of text in images and accept pre-existing transcriptions in CSV and TEI formats (and Word documents converted to TEI). We successfully use an existing crowdsourcing platform to have documents transcribed. Initial preparation is done in a workspace, with TEI as the encoding system. Finalised documents are pushed to a repository for exploration via geographic maps or text searching, and can then be downloaded in various formats for re-use. Once texts are prepared in this way, we submit them to an algorithm to detect non-English items and tag that text as being in the Indigenous languages. https://nyingarn.net.

Keywords*— *Indigenous languages, OCR, crowdsourced transcripts, TEI

I. Introduction

There are over 800 Australian Indigenous languages and many are no longer spoken. For most of them there is very little recorded, and, where there are records, they are often only on paper in a single library. In part, this reflects the destruction of Aboriginal people and societies and the prevailing disregard for Indigenous cultures and languages at the time. There are rare examples of early settlers seeking to understand and record Indigenous languages and this project has built a system called Nyingarn[2] to discover, convert, and make accessible as many of these written sources in Australia's Indigenous languages as possible in a new online digital platform with the text and images of the original documents. There is nothing like this currently available.

It is the responsibility of academics to work with Indigenous people to make these materials available, and to pay attention to any sensitivities there may be with making this information publicly available. These sources are typically wordlists of the language, sometimes a few words, sometimes a few hundred words. The goal is to build a platform in which these manuscript images can be uploaded, transcribed and searched, and to keep adding more manuscripts over time.

We automate as much of the conversion to text as possible with online optical character recognition (OCR) able to deal with most typed text and many handwritten sources. Nyingarn provides specifications for users to add files and that allows for existing transcripts to be added. It also prepares for the time after the current funding is exhausted, ensuring ongoing use of the infrastructure. This, together with a commitment to open formats for the data, will make the content accessible and available for computational treatment in addition to allowing it to be downloaded for re-use in language teaching programs. Nyingarn is directed by thirteen Chief Investigators and by a twelve-member Indigenous Steering Committee. In order to broaden the skill-base of Australian researchers we run training workshops in the methods developed, aimed at students, community members, and more experienced researchers. We are building this platform over three years to allow materials to be located and to allow the platform to become known and trusted and refined following user feedback.

As with many fields, the creation of primary data is essential to doing good research and Nyingarn will facilitate study of, among many other topics: language change over time; uses and range of biological taxa; distribution of songs over time; variation in languages and the diversity of languages recorded close to first contact; relationships between Indigenous people and first settlers; in addition to a range of unexpected topics that will arise by making previously inaccessible material publicly available and searchable. Often, early records for Australian languages provide important information on social life, cultural activities, and Indigenous knowledge. Where languages are no longer spoken daily today, these records can support efforts by Indigenous people to reconnect with heritage, especially in language revival projects that are becoming more common. From a research perspective, each source is more data towards understanding the local language, its history, and its relationship to other languages.

Bruce Pascoe, the noted author and intellectual, discussing the journals of William Thomas, a priceless collection of information about Victorian Aborigines from the late 1830s, notes that they are one of the most important primary sources in Australian history. "So who published his journals? A

[1] The Nyingarn project is funded by the Australian Research Council Linkage Infrastructure, Equipment and Facilities grant LE210100013. Nyingarn has human ethics approval from the University of Melbourne Office of Research Ethics and Integrity (Ref: 2021-22088-20773-6)

[2] Nyingarn is the Nyungar word for echidna. Nyungar is an Aboriginal language from southern Western Australia.

university, a government department, his church, a private researcher? No, the Victorian Aboriginal Corporation for Languages published the papers in 2014." He goes on to note that George Augustus Robinson's diaries, a similarly important set of primary manuscripts, were only published in the 1980s. "While settler reminiscences, football club centenaries and books on outback toilets found plenty of researchers and publishers, two of the most important texts on Aboriginal culture waited over two hundred years." (Pascoe 2014: i-ii) We can characterise this as Pascoe's challenge, and have taken up the challenge by building an online platform to increase the accessibility of primary sources for Australian languages which is a vital step in supporting the growing interest in language and cultural revival.

Our earlier experience in creating the Bates Online project[3] is that Indigenous people generally want to have access to and re-use early sources in their languages. This project made available 22,000 pages of early manuscripts of Australian languages, using the Text Encoding Initiative (TEI) XML to encode links between text and images. This made this set of materials searchable for the first time. Seeing the image of the original document is important for verification of the text, especially where the sources are handwritten and difficult to read. Bates Online is a prototype for Nyingarn but it lacks the ingestion system proposed for Nyingarn, with OCR and transcribing of early sources, and also lacks a generic viewer, allowing new and diverse languages to be included in the platform. Research that derives from historical sources must provide citation to accessible versions of those sources for verification and to assist in transcription.

The fact of materials only existing on paper in a single location makes them difficult if not virtually impossible to consult. Even digital versions of early manuscripts can be difficult to locate, but the Nyingarn platform aims to be a single point of entry to access many documents. Once they are digitised and accessible, new information can be found that was previously inaccessible. A good model for Nyingarn are the platforms for classical European documents, like: Virtual Humanist libraries[4] or Perseus Digital Library[5]. Nyingarn is also inspired by other projects like: TICHA: A Digital Text Explorer For Colonial Zapotec[6]; The Early Modern OCR Project[7]; and Kant's papers[8].

The Perseus project's aim includes "to help make the full record for humanity as intellectually accessible as possible to every human being, providing information adapted to as many linguistic and cultural backgrounds as possible." They have added most of the classical Greek sources and, with the demonstrated success of their platform, have moved into other areas to provide access to many documents, but do not deal with works outside of the classical European canon. The aim is for Nyingarn to provide the same service for Australian language texts that allows addition of new texts over time.

Early written language sources capture information that may not be available to later observers, and can provide information about the societies and natural history in the area where the language was spoken. Making these sources available as text (rather than just as images) similarly allows new information to be explored.

II. Why was this necessary?

When we conceived of this project nothing like this platform existed, and that continues to be the case. There are some examples of single texts or collections of texts online and their presentation includes downloadable pdfs with no text, online text with limited search ability, with a lack of standardised text formats that would allow federated searching. Norman Tindale's collection of many materials from his survey work with Aboriginal people in various parts of Australia is held at the South Australian Museum. The section dealing with vocabularies[9] is given as a catalog with no online access to the materials themselves. Daisy Bates's collection of thousands of pages of notes is held at the Barr-Smith Library and some of these are available as images in pdf documents[10]. The National Library of Australia also holds the same items and does not make them available as digital objects[11]. Individual projects may put manuscript pages online, and google books has pdf versions of some early manuscripts, for example, an 1840 work on a South Australian language is downloadable as OCRed text.[12]

Nyingarn is designed to allow each manuscript to be uploaded, with its transcript, to a workspace, in which it can be improved before it is pushed to a repository that allows searching over all manuscripts. This is extensible, that is, additional manuscripts can be added over time. The current (early 2023) workspace contains 320 manuscripts.

Currently, the Australian government has invested in building a Language Data Commons that will provide access to textual material in as many of the languages (both Indigenous and immigrant) of Australia and its region as possible. Nyingarn is part of this effort and one of our aims is to share infrastructure, in particular, to share metadata schemas, or to provide crosswalks from existing schemas. Further, we are actively building a commons via use of RO-Crate file bundles, described further below.

III. Permission

A critical and founding principle of Nyingarn is that permission must be provided by a relevant language authority before a manuscript can be made public. This recognises our responsibility as non-Indigenous people to work with the speakers of these languages and to create resources in a respectful way with them. At the same time, we are aware that many holding institutions have not made these materials available because they are waiting on permission from as-yet unidentified speakers, and we do not want to replicate this

[3] https://bates.org.au
[4] http://www.bvh.univ-tours.fr/index.htm
[5] http://www.perseus.tufts.edu
[6] https://ticha.haverford.edu/en/
[7] https://texts.earlyprint.org/home.html
[8] http://telota.bbaw.de/kant_op
[9] https://www.samuseum.sa.gov.au/collection/archives/provenances/series/aa338-08
[10] For example, the vocabularies are available here: https://digital.library.adelaide.edu.au/dspace/handle/2440/76574
[11] https://nla.gov.au/nla.obj-229618391/findingaid#nla-obj-359456512
[12] https://rebrand.ly/dmjo5x8

situation with Nyingarn. For material that is in the public domain (e.g. is out of copyright and has been available via services like google books for example) we are considering including the manuscript subject to a takedown policy[13], under which we invite anyone who considers the item should not be made available to make a case to us for its removal.

IV. Planning for the project ending

We have taken seriously the understanding that project endings[14] need to be planned from the start. We know of too many projects that build elaborate infrastructure that is lost as soon as funding ends, and, with it, the critical data that it created or displayed. The design of Nyingarn is based around using Research Object Crate (RO-Crate), a standards-based way of including all metadata together with the files in an item. Each folder thus contains research data together with its description so that it is independent of any catalog database. All files on disk are thus self-describing and can be readily interpreted and re-used without relying on a particular system implementation.

V. The workflow

Manuscript images are accepted as TIFF, and are transcoded on ingestion to jpg and thumbnail versions. If there is no existing transcript, then Nyingarn submits the image files to Amazon's Textract for OCR. The results of this process are exceptionally good for typescripts, and are variable with handwriting, depending on how regular the writing is. Nyingarn is designed to allow source text documents to be created in a number of ways, taking account of existing transcripts, or the preferences of users for a particular transcription system they are familiar with. Systems we know are in use include Transkribus and From the Page, both of which export XML in TEI format. We have an xslt routine that converts each of these on ingest to a common TEI format. For MS Word documents we suggest conversion using OxGarage, that then creates a TEI XML file for import.

An additional source of transcripts is via a crowdsourcing platform, called DigiVol [15] into which we can place manuscripts for typing by volunteers who enjoy helping with the project. Results here are particularly good, but always need to be checked. DigiVol exports text files that are imported into Nyingarn.

Once ingested, the text can be edited in Nyingarn, which presents users with the coded text, marked up with a light TEI encoding, and also a preview view that presents just the text.

All of this preparation of texts is carried out in the Workspace, which presents page images and textual versions to the user for editing. The text has TEI xml tags encoding formatting or structural information. The user can toggle between the plain text version or a preview version that hides the tags, a useful step for less technically inclined users.

The Nyingarn Workspace can be an end in itself as there are no simple equivalent systems available, allowing a user to create well-formed textual transcripts that they can then download and use elsewhere, without passing them through to the repository. Current download options are in TEI, docx, or pdf format, produced via the TEIGarage[16] system.

VI. Technical details

The design (workspace and repository)

The Nyingarn Workspace (the workspace) is an application in which users create items (sets of manuscript images), create arbitrary collections of those items and where they transcribe the textual content of those images or upload pre-existing transcriptions in various formats. They can describe their items using standard metadata terms that are saved as Research Object Crates. The workspace enforces specific naming conventions for the content uploaded in order to produce consistent, well-defined item structures. But otherwise, it allows them to work in the form that best suits their data. So, a user can decide that a small manuscript is modelled as a single item whereas a larger, more complex manuscript is modelled as a set of items brought together as a collection. Indeed the workspace does not limit the grouping constructs so that users can create collections modelling an entire set of manuscripts from the one source or develop other grouping structures as required to best organise the manuscript items.

The technical stack of the workspace consists of a VueJS frontend; a NodeJS API; a PostgreSQL database and a RabbitMQ message queue. The service is built and deployed as a set of docker containers facilitating ease of development on a local machine and simple deployment to the cloud for production. The backend data storage is AWS Simple Storage Service (S3) or a compatible open source service. In the current instance, we are using a compatible service (MinIO) hosted and managed by our partner AIATSIS[17]. The structure of the content on the storage follows the Not OCFL (nocfl) specification invented by one of the authors of this paper (MLR) and is described in a section further on.

The service is specifically designed to support users in the process of uploading manuscript images and transcribing the content as TEI. If a user has a transcription of a manuscript, the service can ingest that transcription (as a TEI file) and make that content available alongside an image of the manuscript for further work. It supports transcriptions in Word format that have been converted to TEI via the TEIGarage (mentioned earlier) web service in addition to transcriptions in CSV format from the DigiVol service.

If a transcription is not available, the service will automatically send the images to Amazon's Textract[18] service for OCR and transcription. The returned content is then formatted as a TEI document and made available to the user to edit and revise as required.

When images are uploaded, they are passed through processing pipelines to produce thumbnails and web safe formats. Specifically, we ask users to provide archival quality

[13] https://www.nsla.org.au/publication/position-statement-takedown

[14] NT participated in the Endings Project based at the University of Victoria (Canada) that explored the need to keep project materials after funding ends: https://endings.uvic.ca/symposium.html

[15] https://digivol.ala.org.au/

[16] https://teigarage.tei-c.org

[17] The Australian Institute of Aboriginal and Torres Strait Islander Studies is a partner in this project and will host Nyingarn when the current project ends

[18] https://aws.amazon.com/textract/

images in TIFF format that the service then uses to create JPG and WEBP representations for the web.

If the content of an image is tabular, users can submit a task via the workspace to send that page to Textract to perform automated table extraction, however, the results of this are variable.

Authentication to the service is via social authentication methods. We currently support OAuth authentication using the Australian Access Federation and Google. We plan to include Microsoft and Meta (Facebook) authentication in the future. In this way we are not responsible for managing user passwords.

The main dashboard upon login shows the user the items and collections they have access to. In Figure 1 we see a listing of the items on the left and a listing of the collections on the right. In the item list we see that the items have varying status. Most items are marked 'In Progress' in blue. When an item is ready to be published it goes into the state 'Awaiting Review' where an admin of the site can verify that the item is ready for publication into the repository. If an item is ready then an administrator can mark the item as "Published" (green) or send it back for further work: "Needs Work" (red).

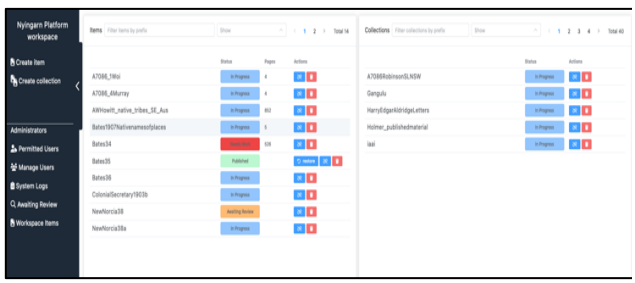

Figure 1. Nyingarn Workspace screen

When a user navigates to an item they see a grid view of the pages of the manuscript (Figure 2). In this example there is a thumbnail of the first 10 images of the manuscript and a basic processing status report (Has a thumbnail been created? Web formats? Has OCR been attempted? Is there a TEI file?). From this view users can delete the whole resource (image and all associated files) if required, perhaps because that image has content that should not have the same permissions in the repository as the remaining manuscript images. Finally, the green background is a visual indicator to the user that they have verified the transcription of that image and marked that resource as complete.

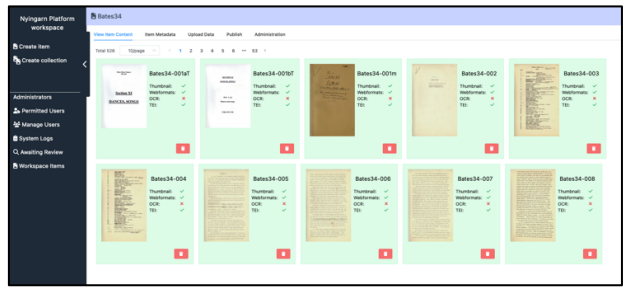

Figure 2. Green shading indicating items have been transcribed

From there they can navigate to a page where they are presented with the image on the left and a transcription editor on the right (Figure 3). In this view the user can collapse the sidebar so their full attention is focussed on the image and the transcription. The image view has controls to zoom and pan the image whilst the transcription editor provides controls to work with the transcription in addition to easily adding TEI XML tags around selections (so the user doesn't need to be a TEI / XML expert).

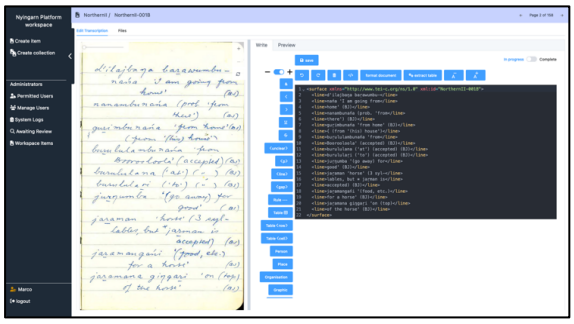

Figure 3. The transcription screen in Nyingarn

The two main transcription import pathways we support are DigiVol (mentioned earlier) transcriptions as CSV files and TEI files (either hand written or produced from Word documents converted by the TEIGarage web service[19]). TEI documents created by hand or by other tools are accepted into the workspace, validated, and then used to create the page transcription stub files. The workspace has code to cleanup and fix many common XML formatting errors and provides a meaningful error report back to the user when it can't ingest a transcription file.

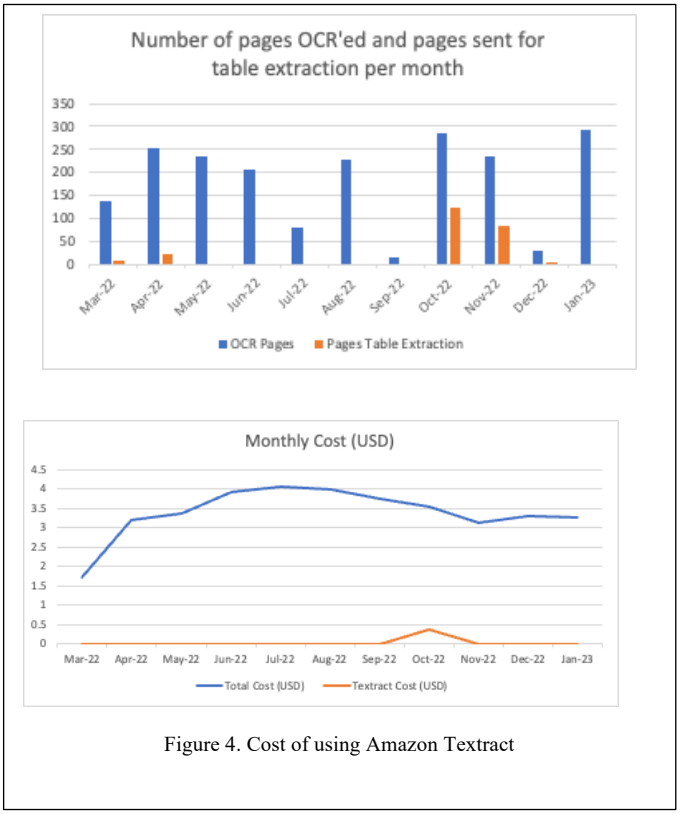

Figure 4. Cost of using Amazon Textract

[19] https://teigarage.tei-c.org/#

VII. Textract Costings

A key part of any project like this is the ongoing cost to run the service. Specifically, using external services like Textract to perform OCR can be a concern as it is hard to estimate the costs up front. Of course this will depend on the exact usage, and Figure 4 shows the costs of operating this service for the last 12 months. In summary the total Amazon spend for our usage in this period (which includes all required services to support Textract) was $37.19USD of which the actual Textract service cost 0.36 USD. Over the time period of March 2022 to January 2023 (inclusive), we processed 1993 documents with Textract OCR and 241 with Textract table extraction.

VIII. Research Object crates and Not OCFL (NOCFL) storage layer

The key design goal that we had for the data storage layer is that the data on disk is stored in a consistent, well described manner and is not dependent on the existence of any service to navigate and understand it. Too often data projects separate the data from the metadata (data on disk, metadata in a database) and if the service fails, they are left with a disk full of data that is not described. Though there are ways to mitigate this (database backups), this model still requires significant effort to 'rebuild' the complete view of the data by joining the on disk representation with the metadata in the database.

Given this, and decades of prior experience operating the PARADISEC catalog [20], we made the decision at the beginning of the project to design our storage layer so that it was a complete representation of the data that could stand alone from the workspace service.

To support this goal we settled on describing our data objects as Research Object Crates (RO Crate)[21]. The RO Crate specification is a method for aggregating and describing research data with associated metadata. The metadata is described using schema.org [22] definitions (mostly, with specific extensions) stored as a JSON linked data file inside the folder containing the data objects themselves.

One of the nice aspects of this specification is that the metadata is written as JSON rather than XML. JSON as a format is more easily worked with in modern software environments than XML. It does not require specific XML skills (XSLT experts) to work with, and is generally readable by users even though it's a machine data interchange format.

For the storage layer, the workspace was initially going to follow the Oxford Common File Layout[23] specification for laying out files on disk. However, previous work using OCFL revealed some unacceptable compromises. The specific issue that we had was that OCFL creates versions at the level of the object. That means any and all changes to the data would result in a version event. In our case we wanted more control of when versioning occurred and that required an approach that supported versioning at the level of the file; not the whole object. (We are not arguing that the specification itself is not sound. If you are building an archival system where you require tracking a path between two file states then OCFL will ensure that you capture that).

One of the authors, MLR, invented the NotOCFL [24] (NOCFL) specification (which is, in effect, a library) to support this usage. The key requirement for our project is that the library needed to work with object storage (AWS S3), allowed versioning at the level of the file, had a simple path structure that was extensible to large repository sizes while still being easy to navigate by administrators using a file browser. We have outlined the differences between OCFL and NOCFL here: https://coedl.github.io/nocfl-js/tutorial-why-not-ocfl.html.

IX. Describo

For the metadata description, the workspace uses the Describo[25] suite of tools and service. The main component is a VueJS plugin that can be embedded into an application to manage the metadata creation for an object. The design is such that the application is responsible for loading / saving the metadata that the component manages internally to itself. The Describo interface is constructed from a profile that is provided to the application and is typically written by a domain expert. It defines the classes and properties that users can describe and the UI is constructed from that definition.
The screenshot shows the Describo tool. In it we can see that there are some groups defined down the left (About, Original Source Information etc) and then properties with data on the right. The layout (groups and properties shown on each tab) are all defined in the profile so an application can adapt its metadata editor exactly as it needs.

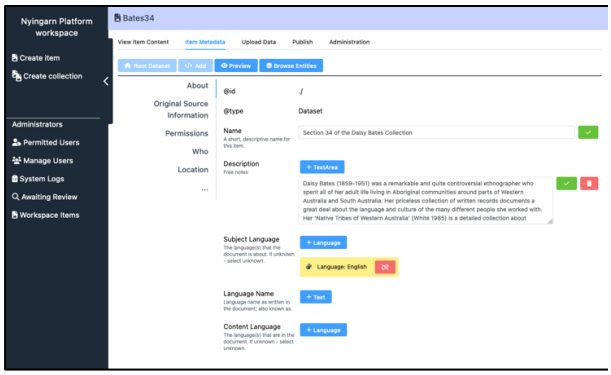

Figure 5. The Describo interface

X. The repository and interface

Once an item has been transcribed, the user can nominate it to be published to the repository. As we are working with Indigenous language materials, we want to include permission from a relevant language authority and that is required before the item can be published. We also require permission from the copyright holder or equivalent. Both of these are provided on forms which accompany items in the repository.

Currently being built, the repository will permit searching over all text, and will include fuzzy searching, allowing for the

[20] https://catalog.paradisec.org.au
[21] https://www.researchobject.org/ro-crate/1.1/introduction.html
[22] https://schema.org
[23] https://ocfl.io. This is an emerging standard and some of the features advocated for in NOCFL may also become available in OCFL.
[24] https://coedl.github.io/nocfl-js
[25] https://describo.github.io

range of variation expected in early writing of Australian Indigenous languages. For example, an Italian observer writes the palatal nasal as ‘gn’, while an English observer may use ‘ny’, or ‘ni’. Each of these options is provided in a fuzzy search, allowing more generous search results than would a literal string. Items can be recalled to the workspace from the repository if necessary for editing or addition of new information.

The repository will be the main search interface to the documents once they have been processed via the workspace.

ACKNOWLEDGMENT

We thank the team of Chief Investigators and the members of our Steering Committee for their help in building Nyingarn.

REFERENCES

[1] Pascoe, Bruce. 2014. Introduction. In Stephens, Marguerita. 2014. *The Journal of William Thomas: Assistant Protector of the Aborigines of Port Phillip & Guardian of the Aborigines of Victoria 1839 to 1843.* Melbourne: Victorian Aboriginal Corporation for Languages. i-ii.