Resources

Making Pacific Languages Discoverable: A Project to Catalog the University of Hawai'i at Mānoa Library Pacific Collection by Indigenous Languages

ELEANOR KLEIBER, ANDREA L BEREZ-KROEKER, MICHAEL CHOPEY, DANIELLE YARBROUGH, AND RYAN SHELBY

The Contemporary Pacific, Volume 30, Number 1, 109–122

Making Pacific Languages Discoverable: A Project to Catalog the University of Hawai'i at Mānoa Library Pacific Collection by Indigenous Languages

Eleanor Kleiber, Andrea L Berez-Kroeker, Michael Chopey, Danielle Yarbrough, and Ryan Shelby

With more than 100,000 books, periodicals, and audiovisual, microfilmed, and rare items, the Pacific Collection at the University of Hawai'i at Mānoa (UHM) Hamilton Library is the premier collection of its kind in the world. Since its comprehensive acquisitions policy was developed in 1969, Pacific Collection librarians have been actively collecting materials—in any language, subject area, reading level, or format—related to all countries and territories of the Pacific. Over the years, the collection has become the world's most diverse resource for comparative regional research in linguistics and languages. An informal 2012 comparative assessment of Pacific-language holdings among the major Pacific library collections in the world (the Australian National University, the National Libraries of Australia and New Zealand, the US Library of Congress, and the University of the South Pacific) showed that in a representative sampling of 100 Pacific languages, the UHM Pacific Collection has more or equal numbers of resources for 72 of these languages. Additionally, the UHM Pacific Collection has the most Pacific-language resources overall, which means it can justifiably be described as the most comprehensive collection in the world for the indigenous languages of the Pacific.

The collection has more than 11,000 volumes written in or relating to the languages of the Pacific. The Pacific region is linguistically quite diverse: The Ethnologue lists over 1,400 distinct languages in Melanesia, Micronesia, and Polynesia (assuming the inclusion of West Papua, Indone-

The Contemporary Pacific, Volume 30, Number 1, 110–122

sia).[1] The Endangered Languages Project shows that nearly all of them are considered to be "endangered" or "severely endangered." The linguistics holdings in the Pacific Collection include dictionaries, grammars, and field notebooks, as well as extensive material written in Pacific languages, such as religious texts, government documents, children's books, and public health brochures. Much of this material is ephemeral and rare and thus may not be found in other libraries in the world. In total, the collection represents almost all the Pacific Islands languages for which materials have been formally or informally published, making it an invaluable resource for the study of Pacific linguistics.

Unfortunately, before the project described in this essay began, we did not know exactly how many and which languages were represented in the Pacific Collection. This is because the library's online catalog did not accurately or adequately describe the items in the collection by language—a problem that was due in part to inadequate standards in common library-cataloging practices (these standards are described in depth in the next section). Until very recently, the potential utility of the collection to a wide range of scholars and language communities had remained untapped. Users seeking materials in Pacific languages could not be confident that they were able to discover relevant materials using Hawai'i Voyager (the UH library catalog); WorldCat (a "union catalog itemizing the collections of 72,000 libraries in 170 countries and territories" [Wikipedia 2017b]); or any other international search engine.

In this essay, we describe our recent three-year project to increase the discoverability and accessibility of the Pacific-language materials in the Pacific Collection by improving and making consistent the descriptive metadata in the catalog, using standards accepted by both library science and linguistic science. Our hope is that our rationale and workflow can be adopted by other area collections around the globe, so that library resource description can increasingly benefit library users who wish to locate materials based on the underrepresented languages they may contain.

Identifying Catalog Problems

Our main goal for this project was to greatly increase the descriptive metadata for all the Pacific-language material within the Pacific Collection by adequately describing the languages represented in the materials. Because the geographic region covered by the collection is among the most linguistically diverse in the world, accurate language information is especially cru-

cial to the usefulness of the collection. We therefore needed to ensure that all Pacific-language material in the collection had adequate and consistent levels of description according to three standards: (1) MARC Code List for Languages,[2] (2) Library of Congress Subject Headings,[3] and (3) International Organization for Standardization (ISO) 639-3 language codes. The first two are considered to be best practices for US libraries, while the third, ISO 639-3, is considered by linguists to be the gold standard for language description. Together, these three standards of descriptive metadata, when applied consistently to bibliographic records in Hamilton Library's online catalog, drastically increase the discoverability of Pacific-language material in the collection.

Before the project began, we identified three major cataloging problems for the more than 11,000 items that are not exclusively written in European languages. The first main problem was that approximately 2,550 non-European-language items lacked Library of Congress Subject Headings or had missing or inaccurate MARC language codes. Due to this lack of basic subject and language description, these items were not linked to similar resources in the catalog and were essentially invisible to researchers.

The second problem, one common to all items, was the use of MARC codes as the primary identifier for languages contained in the items. Indeed, we see this as a problem not just for the Pacific Collection but for all libraries using MARC: even when appropriately applied, MARC language codes often do not adequately describe language to the specificity expected for language research. While the MARC codes are based on the ISO 639 codes for language names, they are correlated to the 639-2 standard, which has since been replaced with ISO 639-3 (as explained in the next paragraph). The MARC language codes are not, in most cases, robust enough to fully describe the rich linguistic content of the collection. This is because for many areas of the world, MARC uses "collective" codes based on language family or geographic region, rather than language-specific codes. For example, the MARC standard subsumes 164 distinct languages under a single collection code, *paa,* for "Papuan (Other)." Specialists seeking to utilize materials in a specific "Papuan" language (such as Auyana, Benabena, or Chuave) are severely underserved by such general descriptions in the catalog.[4] More than 3,500 Pacific-language items in the collection are assigned the MARC collective codes *paa, map* ("Austronesian other"), or *crp, cpe,* and *cpf* (codes for "Creoles and Pidgins (Other)"; "Creoles and Pidgins, English-based"; and "Creoles and Pidgins, French-based," respectively).

The third problem is that more than 700 of the 1,400 languages of the Pacific region have no direct equivalent Library of Congress Subject Heading. These are instead assigned a "complex" subject heading based on geography rather than language. For example, the language Pááfang, spoken in the Hall Islands of Micronesia, receives a complex subject heading "Hall Islands (Micronesia)—Languages," leaving no option for a controlled search based on the language name Pááfang. Another example is Chinese Pidgin English of Nauru, which is given the subject heading "Nauru—Languages"; this example is even more problematic in that nothing in the subject heading indicates that it is a pidgin or what the source languages are.

While the MARC language codes and the Library of Congress Subject Headings are standard practice for library description, linguistically accurate description is confounded by the fact that neither MARC language codes nor the Library of Congress Subject Headings are developed or maintained by language specialists, and problems of omission and commission between library and linguistic standards occur. It is for this reason that the Pacific Collection also had to be described according to a standard generally accepted by professional linguists. ISO 639-3 is the international standard for identifying the world's more than 7,000 languages; it consists of approximately 7,760 unique three-letter codes intended to cover all natural human languages. The list is intended to be as complete as possible and to provide codes for all known spoken, written, living, and extinct languages. ISO 639-3 is curated by SIL International/Ethnologue, which serves as the registration authority and updates the list annually.[5] Examples of language codes include *eng* for English, *aji* for the Ajië language of New Caledonia, and *hwc* for Hawai'i Pidgin.

ISO 639-3 is recognized internationally among linguists, language specialists, and funding organizations as the primary method for unambiguously identifying languages. The name of a language may have alternate spellings (eg, the code *wbp* designates a language of Australia alternately spelled Warlpiri, Wailbri, Walbiri, or Walpiri); a language may be known by alternate names (eg, the code *wos* designates a single language known in Papua New Guinea as Hanga Hundi, Kwasengen, and West Wosera); or a single name may be used for two distinct languages (eg, the name *Kuman* is used for both a language spoken in Simbu Province of Papua New Guinea and one spoken in the Kaberamaido district of Uganda, but the former is given the code *kue* and the latter is given the code *kdi*). Funding agencies like the National Science Foundation require that proposals

include ISO 639-3 codes, and major language-based search engines like the Open Language Archives Community (OLAC) use ISO 639-3 codes as a basis for searching. Adding ISO 639-3 language codes to the Pacific Collection's online catalog allows users to perform much more linguistically accurate searches across the collection.

To remedy these catalog problems, we needed a workflow that would include identifying all the languages present in the Pacific Collection according to ISO 639-3 and adding that code to a dedicated field in each bibliographic record; verifying (when present) or adding (when absent) correct MARC language codes, whether individual or collective; and verifying (when present) or adding (when absent) correct Library of Congress Subject Headings.

Making Pacific Languages Discoverable

In 2013, we applied for funding from the Documenting Endangered Languages Program of the National Endowment for the Humanities to develop the needed workflow. We were awarded $137,317 (NEH PD-50034-14) for three years beginning in August 2014. The principal investigators for this project were Eleanor Kleiber, Pacific Collection librarian, and Andrea Berez-Kroeker, a professor in the UHM Department of Linguistics. Two graduate assistants, who hand-identified the vast majority of languages, were Ryan Shelby (August 2014–June 2016) and Danielle Yarbrough (August 2016–May 2017). Linguists Robert Blust and Piet Lincoln served as language consultants, and Michael Chopey served as the cataloging consultant. We also called on other linguist colleagues worldwide for occasional assistance with language identification. Ellie Kim was a UHM library and information science student who helped enter the improved metadata into the Voyager database.

Defining the Scope of the Project

In defining the scope of the project, we needed to reconcile the geographic scope of the Pacific Collection with the geographic groupings found on the Ethnologue's online presentation of the languages associated with the ISO 639-3 codes. In terms of its acquisition policy, the collection defines the Pacific region as Micronesia, Polynesia, and Melanesia. The Ethnologue, however, defines the Pacific as also including Australia, but excluding West Papua (grouping that region instead with Indonesia). In the end,

Table 1 Publicly Available Online Resources Mentioned

Website	Host	URL
The Ethnologue	SIL International	https://www.ethnologue.com/
The Endangered Languages Project	Alliance for Linguistic Diversity	www.endangeredlanguages.com
Hawai‘i Voyager online catalog	Libraries of the University of Hawai‘i System	https://uhmanoa.lib.hawaii.edu/vwebv/searchBasic?sk=manoa
ISO 639 Language Codes	International Organization for Standardization	https://www.iso.org/iso-639-language-codes.html
ISO 639-3	SIL International	http://www-01.sil.org/iso639-3/
[language map]	ARC Centre of Excellence for the Dynamics of Language	
LCSH: Micronesian Languages	Library of Congress	http://id.loc.gov/authorities/subjects/sh85084873.html
LCSH: Melanesian Languages	Library of Congress	http://id.loc.gov/authorities/subjects/sh85083372.html
LCSH: Polynesian Languages	Library of Congress	http://id.loc.gov/authorities/subjects/sh85104690.html
Making Pacific Language Materials Discoverable	University of Hawai‘i at Mānoa Library	http://guides.library.manoa.hawaii.edu/pacific_language
MARC Code List for Languages	Library of Congress: Network Development and MARC Standards Office	http://www.loc.gov/marc/languages/langhome.html
MARC Standards	Library of Congress: Network Development and MARC Standards Office	http://www.loc.gov/marc/
OLAC Language Resource Catalog	Open Language Archives Community	http://search.language-archives.org/index.html
Pacific Area Language Materials (PALM)	University of Hawai‘i at Mānoa Library	https://scholarspace.manoa.hawaii.edu/handle/10125/42190
WorldCat	Online Computer Library Center, Inc (OCLC)	https://www.worldcat.org/

the items included in our workflow represent just over 1,400 languages of Micronesia, Polynesia, and Melanesia including West Papua and Torres Strait Islands. Note that our classification does not necessarily align with accepted or controversial linguistic families, but this was not a hindrance to our coding procedures because our classifications take place only at the level of language, not of family.

To aid in coding languages, we created a table of correspondences between language names/codes across (1) ISO 639-3 and the associated enhanced information in the Ethnologue for all regions included in the Pacific Collection; (2) the languages that the MARC language code list subsumes under "Austronesian (Other)" or "Papuan (Other)"; and (3) the Library of Congress Subject Headings, specifically the narrower terms listed under the authority headings "Micronesian languages," "Melanesian languages," and "Polynesian languages." (Links to online lists for each of these languages can be found in table 1.) After matching and removing duplicates, we arrived at a list of 1,402 languages.

With a few exceptions, we limited this project to the printed and published items in the Pacific Collection. We did not include audiovisual or microfilm formats, as they are housed in other parts of the library and not physically managed by the collection. Items held in the main Pacific Collection stacks, the collection's rare book vault, and the Pacific creole material held in the Tsuzaki/Reinecke Creole Collection were already more thoroughly cataloged. Items held in "In Cataloging" and "Rapid Cataloging" were less well described, and the latter section had the highest concentration of Pacific-language-only items.[6]

Our Workflow

The first step in the project was to identify which of the more than 100,000 items in the Pacific Collection would be targeted in our project, that is, which items were at least marginally likely to contain Pacific-language text. We started by casting a very wide net as follows before prioritizing the resultant list. The UHM Library's Systems Department generated an initial list of items, and from that Kleiber pulled records for every item for which the catalog already contained a description of a Pacific language, either with a MARC code or with a Library of Congress Subject Heading. This came to approximately 9,550 items.

To this core list, Kleiber then added books that, based on genre, could potentially have contained language data even if they were not already

coded for language. This included material about linguistics, education, anthropology, and biodiversity, as well as items with titles that were clearly in a Pacific language. In total, our "wide net" list of items to be examined over the three years of the project included approximately 17,000 items.

We next prioritized the items on the list. Highest priority was given to items for which the catalog already listed MARC collective codes *map, paa, crp, cpe,* or *cpf*, since we knew that those items contained language data and that we only had to identify these language(s) by their ISO 639-3 code. In other words, these were the items for which our labors would have the highest payoff. Second priority was given to titles that were not yet coded for language but that, based on their subject matter, potentially contained text in Pacific languages—we expected to uncover language data in some but not all of these items. Our third priority was those items that were already coded with one-to-one, rather than collective, MARC codes (eg, Māori or Tuvaluan); these items were assumed to be already well described and could be automatically updated in the library database by Cataloging Librarian Michael Chopey.

The day-to-day workflow of our graduate assistants was facilitated by the development of a custom database interface we dubbed "the widget." Assistants Shelby and Yarbrough used the widget to enter the new language information into the appropriate catalog field, which was then confirmed and committed to Voyager by the library's Cataloging Department.[7]

At the start of each shift, our graduate assistants found a small cart of books that had been pulled from the stacks by an undergraduate student library worker in prioritized order from our master list. Shelby and Yarbrough inspected each book by hand, looking for any materials written in non-European languages. In many cases, the main language of the item was in a Pacific language, but in other cases this language data was only sporadic, for example, glossaries, song lyrics, poetry, chants, and linguistic data sets like word lists.

If an item was found to contain no Pacific-language text, the book was marked in the widget as processed, and it was returned to the stacks. If, however, Pacific-language text was found, the next step was to identify the language and to find the corresponding ISO 639-3 code. In the vast majority of cases this was a fairly simple matter: most materials contained some indication of the language on the title page or in the title, and the assistant needed to only find the corresponding ISO 639-3 code in the Ethnologue and enter it in the widget before returning the book to the stacks.

Other items required a bit more sleuthing to determine the language. When a language was not named outright, or when an alternate language name was used, other context clues in the book, like a village name, a dialect name, or a geographical landmark were often enough for our assistants to determine the language through searches in the Ethnologue, Google, WorldCat, or the Open Language Archives Community. In a few cases, our graduate assistants were unable to identify the language with confidence, and these books were set aside for inspection by linguistic and area experts.

Our final task for the project was to make our updated catalog information more findable by the public. We did this in two ways. First, we asked the librarians in Hamilton Library's Systems Department, which is responsible for programming and maintenance on the Voyager catalog, to add a language-based search function to the catalog.[8] This allows users to search the Pacific Collection with precision by a language's ISO 639-3 code.

Second, our items are now also discoverable through the Open Language Archives Community search engine, thanks to a "crosswalk" we built to convert the Voyager catalog information into OLAC format (Hirt, Simons, and Spanne 2009). OLAC aggregates catalogs of digital language archives worldwide, allowing users to search for language-related resources by, among other things, their ISO 639-3 code. Our items now appear in OLAC searches, as well as in other services that build on the OLAC catalog, like the language map of items in all participating OLAC archives, which was built by the University of Melbourne as part of the Centre of Excellence for the Dynamics of Language (see table 1).

Results

At the time of this writing, the last of the collection's processed items are still being updated in the Voyager catalog, but we anticipate that some 11,500 items will ultimately have been updated by this project, rendering them much more discoverable by language than previously. The UHM Pacific Collection is now also the first library area collection that we know of to include ISO 639-3 codes in its catalog.

Below we provide a few examples of the items that were processed, to give the reader an idea of the kinds of catalog problems that our project corrected.

Example: MARC Collective Code Was Insufficient

Sidney Herbert Ray's *A Comparative Vocabulary of the Dialects of British New Guinea* (1895) contains language data in more than 30 languages. Previously it had been cataloged with the MARC code *paa*. After passing through our workflow, this volume is now cataloged with 31 separate ISO 639-3 language codes, making it more discoverable to researchers interested in particular languages of New Guinea.

Example: Not Previously Coded as Containing a Pacific Language

Prior to the project, thousands of items in the Pacific Collection that contained language data did not include in their catalog description either a MARC language code or an LCSH language descriptor—that is, there was no indication in the catalog that the items contained any Pacific-language data whatsoever. An example is *Niue High School Magazine,* which, on inspection, was found to contain considerable amounts of text in Niuean. It is now coded in the catalog as such.

Example: A Language Has Alternate Names

The Ethnologue and the Library of Congress Subject Headings provide alternate names or dialects for a considerable number of Pacific languages, which means that unless there is a way to link alternate names or dialects together, items in the collection may remain invisible to researchers. An example of this is *A Dictionary of Owa: A Language of the Solomon Islands,* by Greg Mellow (2014). The Ethnologue lists Owa language as having five alternate names (Anganiwai, Anganiwei, Narihua, Santa Anna, Wanoni), which means that a library user searching for one of the alternate names could search on the ISO 639-3 code *stn* and find this volume.

Conclusion

Now that our project to make the UHM Library Pacific Collection discoverable by indigenous languages is coming to a close, we have identified a few potential avenues for future work. First, we believe our project could easily be extended to other libraries with significant holdings of Pacific-language material, such as the National Libraries of Australia and New Zealand, the University of Auckland, and the University of the

South Pacific. Given differing specializations and collection development policies, each of these collections will have unique holdings that were not included in our project. Completing similar projects in each of these institutions would increase overall access to Pacific-language material.

At the start of this project, we had hoped to add new language codes to the MARC list of codes, but because the MARC standards office will only consider adding a new language code once the existence of fifty publications in that language has been demonstrated, we have been unsuccessful. For the vast majority of the 1,402 languages we found in the Pacific Collection, only a small handful of publications exist. However, we have proposed more than one hundred language names to be added to the collective codes, which will at least allow library catalogers to find the appropriate collective code when a language does not have its own MARC code. We expect that the addition of these distinctive language names will help prevent erroneous collective code assignment for Pacific-language materials. In addition, we plan to propose new Library of Congress Subject Headings for the several dozen Pacific languages that do not already have a corresponding, appropriate subject heading.

Finally, the Pacific Islands region is only one of many areas of incredible linguistic diversity—Southeast Asia and Mesoamerica being only two examples of other such areas—and it is our hope that other libraries and collections will seek to implement their own projects to add ISO 639-3 codes to their catalogs. Greater linguistic inclusion among library cataloging practices worldwide will help to equalize the presence of indigenous voices.

* * *

THE AUTHORS ARE GRATEFUL TO *Roxane Gaedeke, Carol Carl, Andrea Nakamura, Ann Rabinko, Wes Poka, and all of the Hawaiian and Pacific Collection student employees for their help in retrieving, reshelving, and otherwise physically processing the language materials. Thanks to the librarians of the* UHM *Hawaiian and Pacific Collections (Stuart Dawrs, Dore Minatodani, Jodie Mattos, and Kapena Shim) and to the* UHM *Department of Linguistics for supporting this project. Thanks to student assistant Ellie Seaton for digitizing the* PALM *materials,*[9] *to Daniel Ishimitsu in the library's Desktop Networking Department for building the* OLAC *crosswalk, and to Arthur Shum and Erin Kim in the library's Systems Department for creating the language code search function. Many thanks to Nackil Sung, head of the library's Acquisitions Department, for building our widget. Gary Simons also provided helpful advice and feedback, for which we are grateful. Special thanks are due to Bill Palmer, Åshild Naess, Alex François, Kaliko*

Trapp, Bob Blust, Piet Lincoln, and Jeff Siegel for their help in identifying some of the more elusive languages in the collection. This work was supported in part by a grant from the National Endowment for the Humanities (PD-50034-14). All errors of omission and commission belong to the authors alone.

Notes

1 The URL for the Ethnologue is listed in table 1, along with all publicly accessible resources mentioned in this article.

2 MARC stands for MAchine Readable Cataloging, and is the standard for representation of bibliographic information in the Library of Congress. For more information on MARC, see table 1 for links to the websites for MARC Standards and the MARC Code List for Languages.

3 "The Library of Congress Subject Headings (LCSH) comprise a thesaurus (in the information science sense, a controlled vocabulary) of subject headings, maintained by the United States Library of Congress, for use in bibliographic records" (Wikipedia 2017a).

4 Another problem of the MARC *paa* code is that the term "Papuan" is understood by specialists to be merely a euphemism for the non-Australian and non-Austronesian languages found in the Pacific, and it does not imply genetic unity. Furthermore, the number of languages that could be referred to by linguists as "Papuan," either geographically or genetically, is somewhere upward of 800—far higher than the 164 languages listed in the MARC code list as falling under the collective heading of *paa*.

5 The URL for this list can be found in table 1, under the title "ISO-639-3."

6 The Tsuzaki/Reinecke Creole Collection is maintained as a separate collection of material within the Pacific Collection stacks. It includes works in pidgin and creole languages from throughout the world and is treated as a closed collection, in that new material is not actively purchased or added to the collection. Newly acquired library materials published in or about Pacific creole languages are still actively acquired as part of the Pacific Collection's overarching acquisitions policy, with these new materials being separately cataloged and added to the main collection holdings.

7 The Cataloging Department at Hamilton Library is separated from the Pacific Collection both administratively (as a separate department) and physically (by five floors). The widget was thus crucial to spanning these separations and communicating catalog updates efficiently.

8 For more on how to search by language code within Voyager, click the "Finding Language Materials" tab on the online library guide Making Language Materials Discoverable (see table 1 for the URL).

9 These materials are freely available via the UHM Library's digital repository, ScholarSpace. See table 1 for the URL.

References

Hirt, Christopher, Gary Simons, and Joan Spanne
2009 Building a MARC-TO-OLAC Crosswalk: Repurposing Library Catalog Data for the Language Resources Community. In *Proceedings of the Joint Conference on Digital Libraries (JCDL '09), 15–19 June 2009, Austin, Texas,* 393. New York: Association for Computing Machinery.

Mellow, Greg
2014 *A Dictionary of Owa: A Language of the Solomon Islands.* Boston: De Gruyter Mouton.

Ray, Sidney Herbert
1895 *A Comparative Vocabulary of the Dialects of British New Guinea.* London: Society for Promoting Christian Knowledge.

Wikipedia
2017a Library of Congress Subject Headings. Wikipedia entry, last updated 24 Feb. https://en.wikipedia.org/wiki/Library_of_Congress_Subject_Headings [accessed 18 Sept 2017]
2017b WorldCat. Wikipedia entry, last updated 11 Sept. https://en.wikipedia.org/wiki/WorldCat [accessed 18 Sept 2017]

Abstract

In this essay, we describe our recent three-year project to increase the discoverability and accessibility of the Pacific-language materials in the Pacific Collection at the University of Hawai'i at Mānoa Hamilton Library by improving and making consistent the descriptive metadata in the catalog, using standards accepted by both library science and linguistic science.

KEYWORDS: library, Pacific Collection, language codes, cataloging, metadata