Digital Scholarship in the Humanities Advance Access published March 8, 2016

What Remains to be Done—Exposing Invisible Collections in the other 7,000 Languages and Why it is a DH Enterprise

Nick Thieberger
School of Languages and Linguistics, University of Melbourne, Australia

Abstract

For most of the world's 7,000 languages, there are few records available via the Internet. Recognizing this digital divide and the consequential underrepresentation of most languages in any linked open data efforts is a motivation for some solutions offered in this article. Efforts to increase the documentation of the world's small languages have led to the development of tools and repositories over the past decade. However, as not all digital language archives currently provide metadata in standard formats, their collections are invisible to aggregated searches. Other repositories (including many institutional repositories—national libraries and archives, mission archives, and so on) have language content that is not noted in the collection's catalog, so is impossible to locate at all via a search based on language names. Finally, there are collections still held by their creators and not in a repository at all, completely hidden from other potential users. This article suggests that it is a digital humanities project to make more information about the world's small languages more freely available, and identifies several means by which this could be accomplished, including a survey to locate more collections; a register to announce their existence; and a documentation index to provide an overview of what is known for each language.

Correspondence:
Nick Thieberger, School of Languages and Linguistics, University of Melbourne, Parkville, VIC 3010, Australia.
E-mail:
thien@unimelb.edu.au

1 Introduction

At a time when digital humanists are concerned with big data and with linked open data, it is worth considering that for most of the world's 7,000 languages (which I will refer to as 'small' languages) there are few records of any kind available via the Internet. In part, this is because few primary records of these languages exist, but, where they do exist, a greater effort could be made to digitize them and to make catalogs of their contents more widely available. They can then participate in what is being offered by emerging methods for searching, annotating, and accessing such material. This article suggests ways in which to make more primary information about the world's languages accessible and observes that the methods align with three defining criteria of digital humanities, namely, 'the concept of performing Humanities research in a distributed digital working environment, which supports equally well: (1) access to the information needed to tackle a research question, (2) the analysis of that information by tools reflecting the methodological requirements of the specific discipline and research problem and (3) the

doi:10.1093/llc/fqw006

publication of the new information gained by the analytical process' (Thaller, 2012, p. 11). The solutions presented in this article fit with the digital humanities emphasis on reusability of research materials (Thieberger, 2014).

This article is written from my experience in working on the Pacific and Regional Archive for Digital Sources in Endangered Cultures[1] (PARADISEC), a digital archive of records of the world's small languages that has been running for 13 years. PARADISEC currently holds 10,400 publicly available items, representing over 900 languages in 5, 700 h of audio recordings.

2 The Problem

The most endangered languages in the world are also the smallest languages, usually in the number of their speakers, but also in terms of resources available in them and about them. Often, like other great cultural treasures, the records of these languages reside in institutions a long way from their source, in the colonial collections of museums, universities, and archives. It is typically impossible for speakers to locate these records in analog form—as papers, audio recordings, or images—let alone to obtain copies of them.

Within the discipline of linguistics, there has been a recent movement to ensure that records are created in better ways than have been used in the past, and that existing legacy records are digitized and made accessible, ideally to their source communities. Under the rubric of *language documentation* (cf Himmelmann, 1998), there is an increased focus on collaborative fieldwork, richer recordings, and a broader notion of what constitutes a record of performances in a language than was the case in the earlier paradigm of *language description* (which documentation does not aim to replace, but to supplement, cf Thieberger, 2014). It also recognizes that the information in primary recordings can always have additional uses beyond those that drove the original researcher. These uses include further scholarly analysis, but it is most likely that speakers and their descendants will be the primary audience.

Two main problems arise, the first is how to locate records of these languages using standard online systems, and the second is how to access analog recordings once they have been located. An ancillary problem that is tied to the solution of the first of the two is knowing what has been recorded for any given language. And, of course, making sense of what is on a recording is a research task resulting from having found what are often legacy recordings with little additional metadata, let alone transcripts.

Over the past decade, a number of digital archives have been established by linguists who recognized the need to ensure that records of small languages are properly curated and that analog media are digitized for preservation. Some of these digital archives have adopted a common metadata system that each archive serves (as an Open Archives Initiative Protocol for Metadata Harvesting feed, https://www.openarchives.org/pmh/) for aggregation via the Open Language Archives Community (OLAC), allowing more targeted language-based searches than can be provided by ordinary Web searches.

OLAC harvests the metadata of fifty-four language archives and aggregates the results every day, providing various visualizations of the results. One view is a page showing what records exist in all of the archives for any given language under specific headings based on the OLAC metadata schema: primary texts, lexical resources, language descriptions, and other resources in the language. It also provides for faceted browsing of the 246,775 records it currently stores. An essential component of this metadata system is the standard language identifier provided by ISO-693-3, a three-letter code that is intended to be unique for each of the world's languages.[2] Adding this simple component to any metadata entry allows it to then interoperate with Web services based on the language identifier. This is a remarkable achievement that is not available to many other research communities and reflects a consensus about the urgency of the work involved in making records of languages that may not be spoken for much longer.

However, not all digital language archives currently provide metadata to OLAC, rendering their collections invisible to the aggregated search. While

their Web pages may be accessible to Web searches, these archives do not allow the targeted search by language that is the focus of OLAC's aggregator. Other repositories (including many institutional repositories—national libraries and archives, mission archives, and so on) have language content that is not noted in the collection's catalog, and the catalog itself may not be available for Web harvesting. Finally, there are collections still held by their creators and not in a repository at all.

This article discusses two approaches to making collections of primary language material locatable and accessible. One is a survey concerned with locating so-called 'hidden collections', and the other is a register to make these and not-so-hidden collections more readily accessible. Both are related in that they aim to build on existing research outputs and to make them available for reuse. While the methods are generalizable to other disciplines, this article describes a register or online catalog of records, specifically of language material, for collections that have no such metadata and for which no other discovery mechanism is foreseeable. Linguists have been able to achieve a consensus on metadata standards that has afforded the kind of aggregation of metadata records not available in other disciplines, but that could serve as a model.

The context for this discussion is the endangerment both of many small languages and also of the records that have already been made of them. While a major effort has been made and continues to be made to conduct fieldwork and to record speakers of these languages today, there is a parallel need to ensure that existing records are also preserved, in particular analog audio recordings. With the scarcity of playback machines and the deterioration of the analog tapes themselves, 'the time window still open for the transfer of dedicated analogue and digital carriers into digital repositories [is] not more than just 20 years' (Schüller, 2008, p. 5).

When discovering a collection of recordings in a particular language, some questions arise: has anyone else made use of this material before, and what else is there in the collective record for this language? By 'collective record' I mean the published material which is relatively easily located, but more particularly the aggregated catalogs of repositories that we know will curate this material safely and make it available. If the language is well-known and has many records, then the urgency of preserving an additional collection is not as great as that of preserving a collection in a language for which there are no other known records. Information about aspects of some of these small languages is found in grammatical descriptions or scholarly articles, sometimes clearly referencing primary records made by the author. It can be difficult to locate these records if they are in the care of the original researcher and the task becomes more difficult after their death. In my experience, most researchers in the pre-Language Documentation era either did not make audio recordings or did not use them after they had served an immediate purpose. Finding those records is the aim of the survey discussed in the next section. Creating a register in which to add these entries so that they are then harvested by OLAC is the topic of Section 3. Allowing the records established by archives to populate a documentation index will provide a current index of what is available for any given language, and that is discussed in Section 4.

3 The Survey—Finding Hidden Collections

Schüller (2008) reports on an Austrian survey of audiovisual collections in 2007 that identified what they call 'hidden collections'. 'This study specifically targeted collections holding primary sources for disciplines like linguistics, ethnography/ folklore, and ethnomusicology, the originals proper of the present day knowledge of linguistic and cultural diversity of Europe and worldwide' (Schüller, 2008, p. 14). The survey had a reply rate of 12%, totaling around 214 responses out of 1,780 questionnaires sent out, and, to their surprise, nearly half of the collections reported were already deposited in an established archive, and only fifty-five of the collections were held at the home of the collector. The report concludes that '[t]he amount of unique research materials, representing primary source materials of the linguistic and cultural heritage of mankind, remains unclear on the basis of the figures available' (Schüller, 2008, p. 5).

I have conducted a preliminary survey of colleagues (linguists, musicologists, and anthropologists) on various listservs, asking them to enter information into a Web-based form. I did not send the form individually to linguists, but relied on lists to distribute the request. The number of responses was predictably low. And, even lower than the proportion of 'hidden collections' surveyed above, only two from fifteen of the collections reported on were already in a repository. However, the responses are still valuable in identifying a number of factors that have prevented the recordings being digitized or deposited in a repository. These results are offered here as part of an ongoing effort to locate such collections, and to find funds to digitize them before they are lost, allowing them to reenter the research programs they were originally created to inform.

The survey questions were kept as simple and easy to answer as possible in order to minimize the effort of responding. The results highlight the lack of responsibility felt by most earlier researchers for the fate of their recordings. They, and the academic disciplines they formed, failed to capitalize on the opportunity they had to ensure that recordings were preserved and made accessible. To be fair, humanities disciplines in general place little or no importance on primary data, unlike the sciences in which verifiability of analysis requires access to the data.

The survey questions were as follows:

(1) Do you know of recordings of small or endangered languages that are not yet digitized? These could be in personal collections or in established repositories that do not plan to digitize their collections. If so, please provide as much detail as you can about the number and type of recordings (reel to reel, cassette, DAT, etc), the content, and the state of their current storage. Can you provide information about who to contact about these collections?
(2) Do you know of collections whose catalogs are not available through federated searches (that is, they are only available if you visit their Web site and not anywhere else on the Web) and for which we could provide a reference to make it easier to find them?
(3) Do you know of repositories of manuscripts that have received little attention from linguists but which are likely, in your opinion, to have linguistic records in them? These may include, for example, missionary archives or state administrative archives.
(4) Please include your name and contact e-mail so we can follow up with you if necessary (e-mail addresses will not be added to any lists).

(Please indicate if you allow us to publish an anonymized version of your response).

The survey form was first publicized among linguistic networks in 2012. It is now nominated as a future activity of the international network of language archives, DELAMAN,[3] which should ensure wider coverage. As a first step, it has revealed an interesting variety of collections, each with characteristics that are significant for the effort of making such collections available. At a time when funding for digitization is difficult to obtain, it is important to recognize that unique cultural heritage recordings such as these are at risk of being lost. A summary of nine responses and an observation about the broader significance of each is given below.

(1) Twenty-two tapes of an indigenous Sudanese language are held in Washington, DC by a retired linguist—how can they be digitized to suitable standards and where should they be stored? Twenty-two tapes are a manageable number and should not cost too much to digitize (current estimates would be in the order of US $2,000, including metadata entry, tape cleaning, and so on). There are also a large number of notes that need to be scanned. For a retired researcher, it may not be easy to access the equipment needed to do this work.
(2) Several hundred cassettes in Solomon Islands language, particularly valuable as they are recorded by a speaker, so capturing lots of natural speech. Digitizing such a collection is a serious undertaking needing significant funds.
(3) The tapes of Papuan language are in Stockholm, stored in a box, but the recorder is based in Chicago and is still an active academic.

(4) Colorado, USA, has a dozen reel-to-reel and two dozen cassette tapes of various African indigenous languages with a senior linguist concerned to make the collection safe and not being sure what to do. There is no suitable archive locally, nor source of relevant advice.
(5) Tapes were deposited with a national cultural center in a small Pacific country that may or may not have the resources to look after them. It does not publish its catalog (if it actually has one), and so it is not clear if these tapes need to be digitized or not, or what conditions may be placed on access to them.
(6) A recent MA in Linguistics at one of the PARADISEC consortium universities has tapes stored in boxes. Paper transcripts may have been thrown out. It shows lack of communication even within our own departments.
(7) It was reported that the Parry Collection of Harvard University has a set of tapes of the Kesar epic in the Purik dialect of Ladakhi (Tibet) and has no plans to digitize them. However, a search of the Parry collection's catalog does not find 'Ladakhi'.
(8) "A small collection of ten cassettes for Hoan was available (I digitized them myself, but I do not know if the quality is as good as it could be. I just played from a cassette into my Marantz 660PMD)". This raises the problem of methods used in digitization. Without understanding the need to assess cassettes before playback, the process described here may damage the tape and will certainly not result in the best digital file. This points to the need to provide training in better methods and to provide accessible digitization systems.
(9) The Humboldt University in Berlin holds recordings made of foreigners held in a Prisoner of War camp at Spandau, Berlin, and other Prisoner of War camps in 1915. There is a listing of their Polynesian holdings[4] but there are others as well, as the foreigners recorded came from many places, including a shipload of South African seamen from various tribes. While these recordings are not in danger, they are inaccessible as there is no current archivist in charge of them. This is a digital collection held in a repository, using idiosyncratic language names (see Fig. 1) that obscure what the actual content of the collection is.

Using the information from this survey, we have worked with the collectors to accession those recordings that we can. The Papuan tapes are likely to be digitized with existing funds. We have applied for and received a grant from the Endangered Archives Programme and have digitized and accessioned 200 tapes from a collection held by the Solomon Islands Museum. Several other collections identified are also now being digitized and archived.

The main conclusions that can be reached after summarizing the survey results are as follows: there is clearly a need for training in the creation, description, and archiving of research records; in order to achieve this as efficiently as possible there is a need for simple metadata entry tools; there are not enough archives, and so we need more standards-compliant repositories; and, in order to motivate the recalcitrant, there is a need for recognition of collections of primary material as a form of scholarly output.

4 The Register of Language Identifiers

The relationship of a person to their language becomes increasingly significant as the number of speakers declines, that is, the value of language records for languages with few or no speakers can rise inversely in proportion to the number of speakers for whom it is an ancestral language.[5] The extraordinary value of such records suggests that announcing their discovery, if they were previously unknown to specialists in the area, should be a priority for researchers and for speakers. What I mean by discovery is, as discussed earlier, locating records unknown to the researcher in a location that has either not included language codes in its catalog or else has no catalog of its collection. For example, transcripts of court hearings in the local language, a biologist's collection in a museum that has local

Fig. 1 A catalog entry (http://www.sammlungen.hu-berlin.de/dokumente/11905/) from the Humboldt Museum, Berlin, with Tahitian erroneously listed as an Australian language in the header, and later as a Polynesian language. There is no ISO-639-3 identifier provided for the language

names for each item, or surveyor's notebooks that include vocabulary from local languages. Each of these could be located by a researcher in the course of their work, but how can they then announce the language content in a way that would be located by others? What form could such an announcement take?

The survey results, together with our experience over a decade of running PARADISEC (Thieberger and Barwick, 2012), suggest the need for a systematic register of collections that includes language identifiers, in particular ISO 639-3, which is then harvested OLAC. A three-letter code is assigned uniquely to each language, avoiding the computationally fraught problem of multiple spellings or names of languages. For example, if the language name is also a common word in another language (e.g. *Noone*, *Karen*, *Kola*, *Titan*, *Maria*, *Mono*, *Mum*—which are all language names) then searching for them by that name will result in too many hits to be useful. Using ISO 639-3 can make such searches more successful, so the language *Kola*,

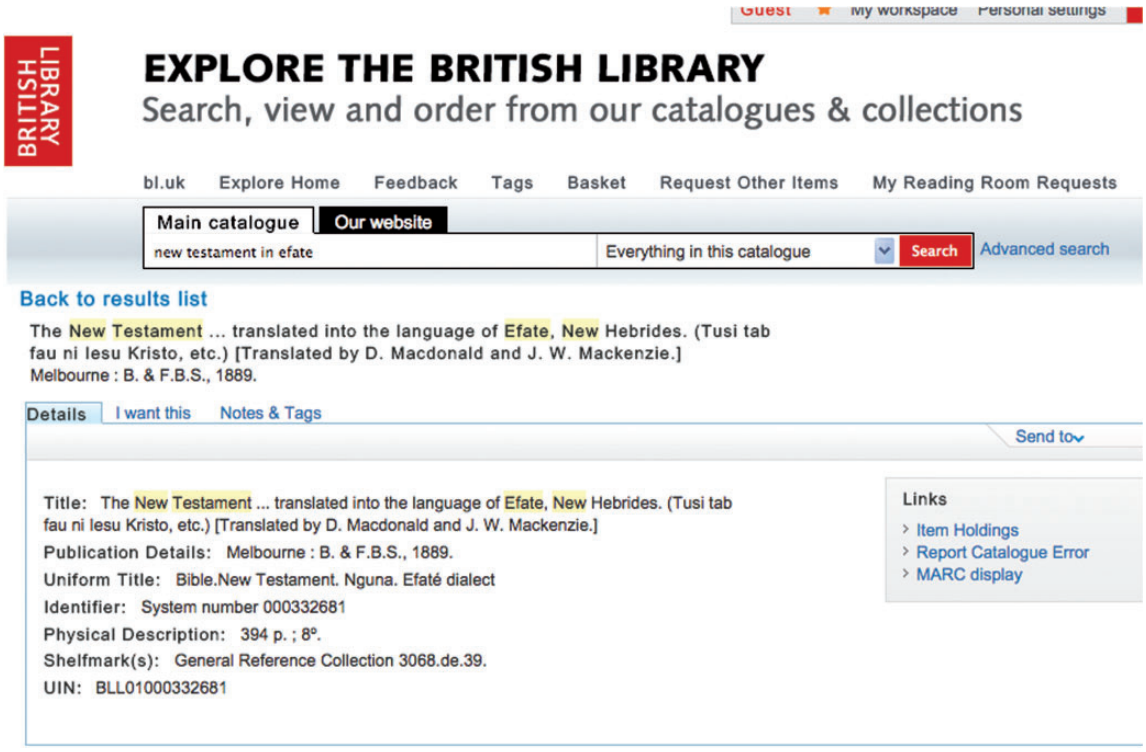

Fig. 2 Catalog entry in the British Library showing no language subject

mentioned above, has the code 'kvv', and *Maria* (in Papua New Guinea) has the code 'mds'. If you can find a way of associating your newly found collection with this standard code (e.g. by lodging records in an archive), then searches will be more targeted. There is room for much improvement in the mechanism for identifying languages, and there will always be sociopolitical disagreements about how to split or group language varieties (Morey *et al.*, 2013) but, in the main, the use of these codes provides better results than does their absence.

As seen in Fig. 2, incorrect language attribution can make it difficult to locate items in a catalog, but the lack of a language identifier altogether is also problematic. Fig. 2 shows a catalog entry from the British Library for a work that is entirely in the South Efate language of Vanuatu (ISO-693-3 code 'erk') and Fig. 3 shows the same item in the National Library of Australia, with the Subject 'Efate language' and the language identifier 'Austronesian (Other)'. In neither case would a search for the standard name 'South Efate' or 'Efate, South' find this item, and in neither case does this item appear in OLAC's resource list for this language.

The register proposed here provides metadata in an open archives initiative-compliant form, allowing records to be found in generic language searches.[6] Not all repositories can provide metadata using ISO-639-3 language codes, so it is useful to provide a mechanism whereby researchers can build this resource as they discover new material. It would make sense to gamify the entry of items into the register, perhaps rewarding those who have supplied information with honorable mentions. The register would also have to be publicized widely to ensure it is taken up.

In general, repositories are just unaware of standards rather than being reluctant to share data, hence the need for them to either change their metadata system (which is unlikely) or for a register of the kind described here, that points to their collections.

Title	Tusi tab fau ni Iesu Kristo nauot anigita nag i muti gita : = The New Testament of our Lord and Saviour Jesus Christ / nafisan ni Efate, New Hebrides = translated into the language of Efate, New Hebrides.
Uniform Title	Bible. N.T. Efate.
Also Titled	New Testament of our Lord and Saviour Jesus Christ translated into the language of Efate, New Hebrides
Other Authors	British and Foreign Bible Society. Victoria Auxiliary
Published	Melbourne : Printed for the British and Foreign Bible Society by the Victoria Auxiliary, 1889..
Physical Description	394 p. ; 19 cm.
Subjects	Efate language -- Texts.
Language	Austronesian (Other)
Dewey Number	220.5994
Libraries Australia ID	24155203
Contributed by	Libraries Australia

Fig. 3 Catalog entry in the National Library of Australia showing a broad language subject but no standard language term

Table 1 A sample listing of references in the PARADISEC catalog to external items

PARADISEC identifier	Title
External-Tikopia	Tikopia, The Liturgy in Tikopian
External-Vaturanga	Vaturanga, Christian material in Vaturanga
External-Aoba	Aoba, Portions of the Book of Common Prayer in Aoba
External-Arosi	Arosi, Prayer Books in the Arosi Language
External-Bauro	Bauro, Portions of the Book of Common Prayer in Bauro
External-Biak	Biak, On line language documentation for Biak (Austronesian)
External-Melanesia1938	Melanesia, The 1938 Book of Common Prayer for Melanesia
External-VARISI	VARISI, Varisi and Baniata language material
External-Lakona	Lakona, Morning Prayer, Evening Prayer, Prayers and Thanksgivings, Litany, Eucharistic Prayers and Catechism in the Lakona language.

The register could simply assign a language code to the URI of the collection and then serve that information for OLAC to harvest. The Rosetta project has established this kind of register of items in the Internet Archive (discussed below), and the same model could be applied more generally to any source located in a repository that does not identify languages using standard codes. The pilot version we have established works by inserting records into PARADISEC's catalog. So, for example, there is a Web site produced by the Anglican church that contains textual versions of early missionary translations in languages of the Pacific. This rich collection of unique transcriptions of early records has no language identifiers, and so creating a record in our catalog makes it visible to OLAC's search.

To avoid the problem of broken links, and recognizing that the pages are unlikely to change once published, we point to the most recent version of the site in the Internet Archive. We currently point to some thirty such collections; few are listed in Table 1. Given the success already achieved by OLAC, and with the support of DELAMAN, such a register of language content of collections should continue to grow and to be a useful resource. These collections can be considered to be 'low hanging fruit', that is, they are already digitized and online, and only need to have a pointer to make them enter into the larger listing of information in that language, and, eventually, to provide access to the primary data in that language. With support from the community of linguists we plan to build a stand-alone register in 2016.

Table 2 Comparison of current Web sites listing information about the world's languages

Current websites listing the world's languages	Coverage —all languages?	Listing resources	Links to primary sources	Users can add information	Cites OLAC	Provides metadata for OLAC harvest	Hosts primary material
Proposed documentation index	+	+	+	+	+	+	−
OLAC[a]	+	+	+	−	+	−	−
Virtual language observatory[b]	+	+	+ via OLAC	−	+	−	−
UNESCO[c]	− (2471 languages)	−	−	−	−	−	−
Rosetta[d]	+	+	+	−	+	+	+
Ethnologue[e]	+	+	−/+	−	+	+	−
Endangered languages site	− (3231 languages)	+	−/+	+	+	−	−/+
Project Joshua[f]	+	−	−	+	−	−	+
Wikipedia[g]	+	−/+	−	+	−	−	−

[a]www.language-archives.org
[b]http://catalog.clarin.eu/vlo
[c]UNESCO Atlas of the World's Languages in Danger: http://www.unesco.org/culture/languages-atlas/
[d]http://rosettaproject.org/projects/rosetta-platform/
[e]http://www.ethnologue.com/
[f]http://joshuaproject.net/
[g]https://en.wikipedia.org

As is clear from Table 2, most current sites do not feed new information to aggregators, in general they are clients of OLAC, so any new information fed into OLAC by the register would then be made available via each of these other services.

In the first envisaged iteration of this register, we will use a version of our catalog (NABU[7]), with the functionality it already provides and its existing feed to OLAC. We will work with DELAMAN to publicize the register.

NABU is itself only as enduring as the funding to support it and the management of its servers. While it has lasted 13 years so far, we are aware that it will, at some point, be likely to be incorporated into other systems. To avoid the worst outcome of the loss of the catalog we write, each catalog entry to an XML file every time is saved. This means that each item in the collection is stored together with an XML description of its contents.

5 The Language Documentation Index

The last piece of infrastructure discussed here is the documentation index. Having established a register of primary material held in collections, it is then possible to use this information to build an automated index of what is known about each of the world's languages. Such an index serves several purposes. First, it is a good public information tool, showing a general user how little we know about the world's languages. Second, it serves two useful research functions, one is the obvious point of reference to find out what is known for any given language or geographic area, and the other is a kind of reward for scholars who have entered their information into repositories that feed the index. This index should be distinguished from endangerment indices (of which there are several existing examples, see e.g. Harmon and Loh, 2010), which measure the relative vitality of the language and not the amount of documentation available for it.

An early version of a documentation index was presented by Wurm (1963, p. 137), who set out a scale that combines features of language vitality (number of speakers and degree of endangerment) with a four-point scale for amount of vocabulary recorded, and a five-point scale for amount of morphosyntax recorded.

There are a few sources that give such an index for particular regions. McConvell and Thieberger (2001) implemented an index for Australian

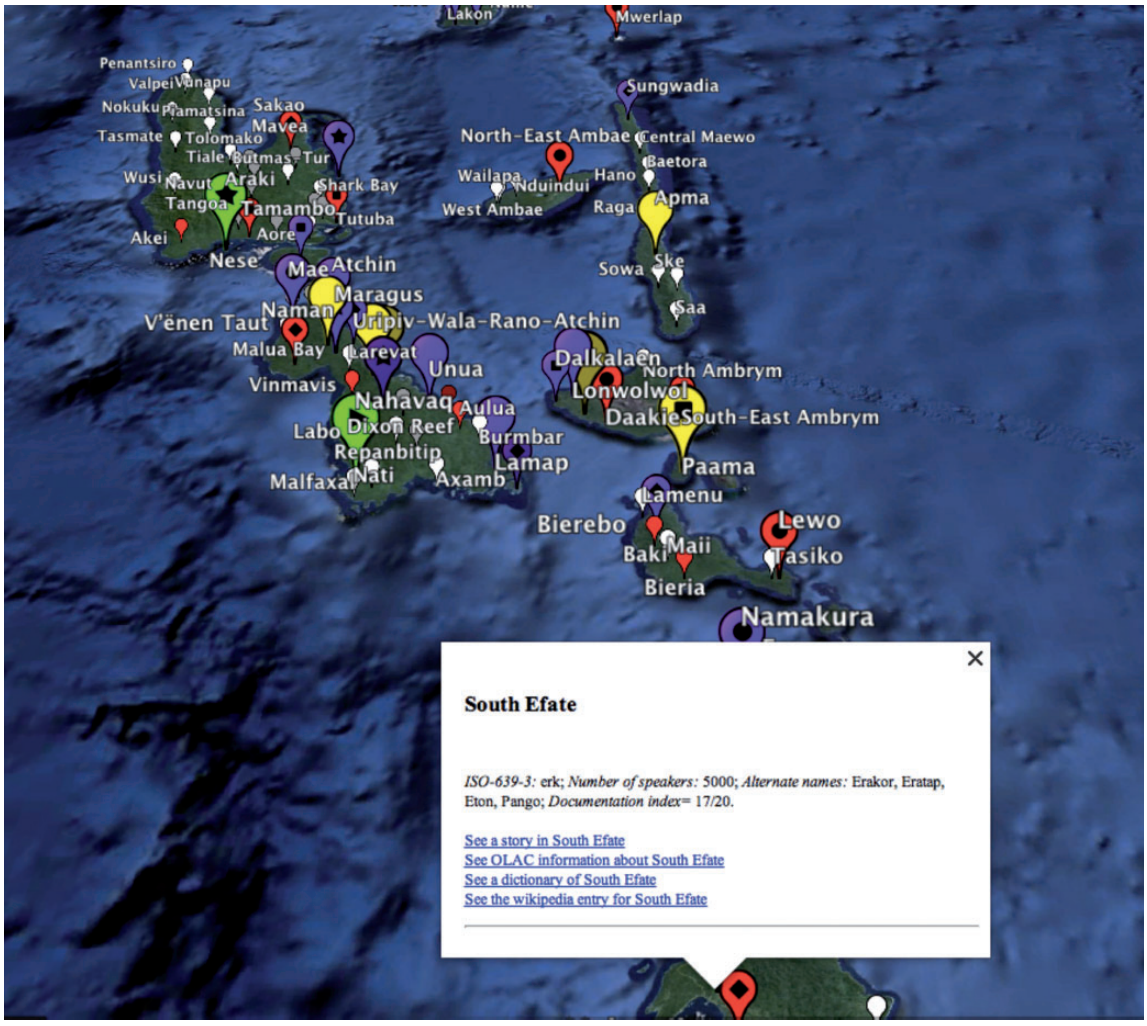

Fig. 4 Visualization of a documentation index for Vanuatu languages, where color and size of icons indicate degree of documentation for a language. The very small white icons represent virtually nothing known about a language.

languages and this was subsequently built into Austlang,[8] a directory of information about Australian languages that assigns up to four points, depending on the amount and quality of each of four features—word list; text collection; grammar; and audiovisual—and the score gives indicative figures for the language.

Another example is Lynch and Crowley (2001, pp. 17–19) who provide a five-star system for languages of Vanuatu. A subsequent survey of Vanuatu languages (Thieberger, 2013) used a 21-point scale assigning 1–5 points for each of four categories: grammar; lexicon; texts; and media corpus.

The resulting index creates a map[9] showing markers varying in size reflecting the documentation index for that language, see Fig. 4. Such a map should be constructed automatically from Web services such as, for example, the OLAC. This is the model proposed for a documentation index of the world's languages.

In any such documentation index, it is important to distinguish cases in which it is unknown what resources may be available from cases in which it is known that there is no information, that is, a zero score needs to be distinguished from a 'lack of information' score.

A crude measure could simply list the number of items (dictionary, grammar, texts, recordings, etc) for a given language, but that would take no account for quality of the items counted. For example, a thousand audio snippets of words in a language would be weighted higher than a few audio files with the same content, or a complex dictionary could be outweighed by a number of lexical files with simple word-for-word correspondences. So it is likely that the documentation index requires human judgments of the quality of the documentation. Given the scale of the information to be dealt with it may make sense to apply an automated metric in the first instance and to indicate to users when an index rating is automatically arrived at rather than having been considered by a local expert, perhaps providing two views, one of the automated index and the other of the human-created index. The range of five points can take into account quality as well as issues like the range of text types, for example, and the degree of diversity of speakers included in the media recordings. Such judgments can be facilitated by having a page of aggregated material available from which to assess the material for any given language. A map that is automatically generated from a feed from archives then gives an immediate visual impression of where more work is needed, as can be seen in Fig. 4.

Why is there a need for such a documentation index given the number of sources of information on small languages? Most existing language directory sites do not contain primary material themselves but in general summarize each other, usually based on OLAC's aggregation of information in the fifty-four language archives it currently harvests from, as summarized in Table 2. Both UNESCO and ELCat are thin in that they contain little or no primary material and they only deal with what they define as endangered languages. As can be seen, only Rosetta provides primary material (including, as already mentioned, providing a useful service of indexing the Internet archive[10]), while the other sites are aggregators of existing information. Two other sites with potentially relevant information (LLMap[11] and the World Oral History Project[12]) were not working at the time of writing this article.

6 Conclusion

The lack of care taken by scholars to ensure the survival and accessibility of their research records reflects a culture that emphasizes analysis and research results without requiring verifiability of assertions with reference to the primary data on which they are based. This failing by our forebears has resulted in a number of orphaned collections of research recordings that now need our urgent attention. A survey of these collections aims not only to identify where effort needs to be focussed, it also raises a more general awareness about the urgency of preserving analog collections. The online registration of orphaned items facilitates their reuse and allows them to be discovered by others. Having increased the coverage of our automated aggregator we can then use its outputs to build a dynamic documentation index. In this article I have shown a model for the location of primary materials relevant to linguistics in which a major standardization issue has been agreed to by a number of researchers—the use of standard language identifiers as the lynchpin that allows all of this to work.

Much remains to be done to extend the reach of digital language archives, including assisting in locating legacy collections, describing and digitizing them, connecting with source communities/individuals, creating a means for online annotation (crowdsourcing), and of valuing the collections (both monetarily and academically). The three projects described in this article go some way to building a research base for all of the world's languages.

Acknowledgements

Thanks to two anonymous reviewers for comments that have improved this article. I acknowledge the support of the Alexander von Humboldt Foundation

and the Linguistics Department of the University of Cologne who hosted me in 2013 and 2014.

Funding

The work reported on here was supported by Australian Research Council grants DP0450342, DP0984419, and FT140100214.

References

Harmon, D. and Loh, J. (2010). The index of linguistic diversity: a new quantitative measure of trends in the status of the world's languages. *Language Documentation and Conservation*, 4: 97–151.

Himmelmann, N. P. (1998). Documentary and descriptive linguistics. *Linguistics*, 36**:** 161–95.

Lynch, J. and Crowley, T. (2001). *Languages of Vanuatu: A New Survey and Bibliography*. Canberra: Pacific Linguistics.

McConvell, P. and Thieberger, N. (2001). *State of Indigenous Languages in Australia – 2001*. Australia State of the Environment Second Technical Paper Series (Natural and Cultural Heritage). Canberra: Department of the Environment and Heritage. http://repository.unimelb.edu.au/10187/485

Morey, S., Post, M. W. and Friedman, V. A. (2013). The language codes of ISO 639: A premature, ultimately unobtainable, and possibly damaging standardization. *Paper presented at 'Research, records and responsibility: Ten years of the Pacific and Regional Archive for Digital Sources in Endangered Cultures.'* http://hdl.handle.net/2123/9838

Schüller, D. (2008). *Audiovisual Research Collections and their Preservation*. Amsterdam: European Commission on Preservation and Access.

Thaller, M. (ed.), (2012). Controversies around the digital humanities. *Historical Social Research*, 37(3).

Thieberger, N. (2013). Language Archives for the Pacific. *Presentation at the DoBeS conference Language Documentation: Past – Present – Future*, Hanover, June 5–7, 2013.

Thieberger, N. (2014). Digital humanities and language documentation. In Gawne, L. and Vaughan, J. (eds), *Selected Papers from the 44th Conference of the Australian Linguistic Society, 2013*. Melbourne: University of Melbourne, pp. 144–59. http://hdl.handle.net/11343/40961

Thieberger, N. and Barwick, L. (2012). Keeping records of language diversity in Melanesia, the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC). In Evans, N and Klamer, M (eds), *Melanesian Languages on the Edge of Asia: Challenges for the 21st Century*. LD&C Special Publication No. 5. Honolulu: University of Hawai'i Press, pp. 239–53.

Wurm, S. A. (1963). Aboriginal languages. In Australian Aboriginal Studies; Stanner, W. E. H. and Sheils, H. (eds). *Australian Aboriginal Studies. A Symposium of Papers presented at the 1961 Research Conference*. Melbourne: OUP, pp.125–65.

Notes

1 http://paradisec.org.au
2 http://en.wikipedia.org/wiki/ISO_639-3
3 DELAMAN is the Digital Endangered Languages and Musics Archives Network. The survey is at http://www.delaman.org/project-lost-found/
4 http://www.sammlungen.hu-berlin.de/schlagworte/6568/dokumente/
5 I am indebted to David Nash (personal communication) for this observation.
6 For example, this record in the PARADISEC catalog http://www.language-archives.org/item/oai:paradisec.org.au:JL1-link refers to recordings in the Smithsonian in the Kilivila language (ISO-639-3 code kij) and can be found in this aggregated page of material about Kilivila: http://www.language-archives.org/language/kij
7 NABU is the name of the catalog which can be seen at http://catalog.paradisec.org.au, and the source code at https://github.com/nabu-catalog
8 http://austlang.aiatsis.gov.au/
9 http://www.paradisec.org.au/blog/2014/06/language-documentation-index/
10 https://archive.org/browse.php?field=subject&mediatype=texts&collection=rosettaproject
11 http://llmap.org/
12 http://oralliterature.org/dadabik2/dadabik_4.2/program_files/index.php?function=show_search_form&table_name=db5