PARADISEC and ISO-639-3
Nick Thieberger

[Notes to accompany the slide presentation]

Abstract

As recently argued by Kamusella (2012), the regime of language codes is problematic for a number of reasons ranging from the very definition of 'language' to the contingent nature of language naming. While recognising that there are many difficulties associated with establishing standard codes, I suggest that we still need some mechanism for identifying what resources (recordings, texts, dictionaries, analyses and so on) are available for particular languages, especially based on my experience with the Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) which has used ISO-639-3 codes since its catalog began in 2003. In the main, these codes have served a useful function and are uncontroversial in identifying the language variety recorded in the collection item. In the PARADISEC catalog we have always had an additional field for 'Language as given' and there have been instances where we have used a code for a nearby language when there was none for the named variety included in the collection. We also provide a geographic reference for the language, either as a bounding box drawn by the depositor or else derived from the ISO-639-3 code (using centroid data bought from GMI[1]). Each of these metadata elements facilitates locating the item in the collection and that is its main function. In this talk I will summarise some of Kamusella's arguments with reference to the experience of building the PARADISEC collection of linguistic material. I will explore ways of using standards that do not necessarily reify language varieties

Introduction

As publishers of information about collections of language material, we want to maximise the locatability of the material and, at PARADISEC, we have found ISO-639-3 to be a useful standard to include in among a suite of metadata items. I must stress that I am talking from the perspective of an archive that needs to establish the relationship between language identifiers and a piece of work, that is, a recording, or text that is in or about the language in question. The correspondence between the name given in legacy materials (like tape boxes, fieldnotes) works for much of the legacy material we are digitising, mainly, I suspect, because the linguists who created these recordings also provided the information that Ethnologue is based on and so there is a match between the varieties named in their metadata and ISO-639-3.

On the other hand it is clear that the language names provided from more recent fieldwork are not always as amenable to standardisation in ISO-639-3. This is particularly the case when fieldwork is conducted in a village that is not the centre (the mission base or fieldwork location, cf Kamusella 2012: 74) from which the original language designation was determined.

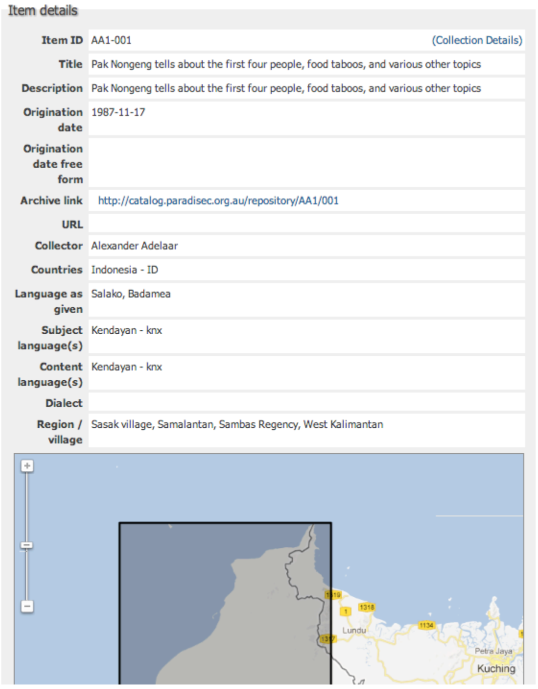
How the language codes came to carve up what we could call the linguasphere is a topic of Kamusella's article and need not be repeated here. Despite his critique, Kamusella nevertheless acknowledges the need for language name standards (which

[1] http://www.worldgeodatasets.com/language

he calls (ibid: 59), I think inappropriately, ‘language standards’), as he says “For better or worse, the emerging global regime of language recognition is what it is” (78) and “it has become an indispensible element of cyberspace”.

What I want to focus on is how to use the codes as a finding aid for material stored in research collections. Given that ISO-639-3 exists and has some infrastructure associated with it, we (PARADISEC) decided to include it in our first metadata design. We have to weigh up the benefit of using some standard to identify content in collections against a free-text search for non-standardised terms associated with a language variety. While not all archives agree with the use of standard language codes we have chosen to allow the use of Ethnologue’s standard codes as one of a set of descriptors that narrow a search for language material in the collection. The other descriptors include: a ‘language as given’ field in which the depositor can enter the name of the language as it is used in their research or by the people they work with; a geographic bounding box that shows where the language was recorded; a ‘region/village’ field; and a ‘dialect’ field.

So, for item AA1-001 in our collection, the closest ISO-639-3 code is for Kandayan, but the language as given is Salako and Badamea. Village names are also given, as is geographic information via a DC:bounding box of min/max lat/long. Each of these pieces of information provides searchable cues to locating the material described.



Item details

Item ID	AA1-001 (Collection Details)
Title	Pak Nongeng tells about the first four people, food taboos, and various other topics
Description	Pak Nongeng tells about the first four people, food taboos, and various other topics
Origination date	1987-11-17
Origination date free form	
Archive link	http://catalog.paradisec.org.au/repository/AA1/001
URL	
Collector	Alexander Adelaar
Countries	Indonesia - ID
Language as given	Salako, Badamea
Subject language(s)	Kendayan - knx
Content language(s)	Kendayan - knx
Dialect	
Region / village	Sasak village, Samalantan, Sambas Regency, West Kalimantan

Given that we clearly do need some way of identifying languages, while at the same time recognising that the names or codes are a heuristic that should not be taken as a perfect reflection of reality, what alternative to using ISO-639-3 is there? The first alternative could be to decide that no standard terms are necessary and that any language name can be used since search engines will find it. This was the topic of a discussion on our blog in May 2011[2] with another linguist talking about what kind of

[2] see the blog discussion of these issues here: http://www.paradisec.org.au/blog/2011/05/you-gotta-be-in-it-to-win-it/

metadata terms to mandate in our catalogs. I noted that my interlocutor “dismisses efforts to standardise terms as being outdated (in the olden days it seems, ‘key metadata notions were interoperability, standardisation, discovery, and access’) and advocates a relativist metadata mush in which there is a ‘focus on expressivity and individuality in metadata descriptions’. Expressivity and individuality certainly have their place, but they don’t help when it comes to targeted location of information, especially at the scale of material to be searched on the web.

While the conversation covered more than just language identifiers, I think the principle is the same. A product of allowing users to enter their own terms rather than providing them with a set list and a freeform field for their own version is that a collection will not have any standard terms for locating information. Without standard terms, looking for ‘songs’ will not find song, looking for ‘kastom’ will not find ‘Custom description’ or ‘Custom narrative’ or ‘Custom story’, let alone ‘Folk Tale’, ‘Narrative’, ‘Myth narrative’, ‘Narration’, ‘Narrative from visual prompt’, and many more. Who knows what ‘Chronicle’ or ‘Semi-spontaneous interview’ will find. And it is nice that the terms can be in any language, but that reduces the predictability of the search finding anything even more. I can’t see why it is an advantage to have all of those words rather than a standard set of terms and then a free form field in which such stream of consciousness tags can also be listed.

Thus, for example, the Arandic songs project is tagged with ‘Language: Arandic’, [NEXT] which is not part of the standard language terms lists. Searching for the more usual term ‘Arrernte’ [NEXT] does not locate these items in the first ten pages of a google search (I gave up looking any deeper than that). The term ‘Arandic’[NEXT] does feature in the more inclusive lists like Glottolog and Multitree.

By participating in international standards, the items in a collection could be found by [NEXT] pages like this: http://www.language-archives.org/language/are

Here, the standard three-letter code at the end of the URL links to a page listing all available information held in participating archives, and this is updated periodically, effectively providing a dynamic documentation index. Of course there are still problems with the three-letter codes, but they are improving over time, and this and other issues could be improved by cooperation rather than competition from the small community who are doing this work.”

The need to identify names of language varieties recorded in existing documents has been addressed in several projects.

The Library of Congress MARC system’s language authority files are very limited (515 items[3] listed – [NEXT] including various levels of grouping, and [NEXT] Klingon) and, helpfully, also use a (different) three-letter code [NEXT] as seen in the middle of this image, but with no apparent process for change except that someone requests it by appealing to a published work. [NEXT] So, for example the librarian at the University of Hawai’i requested the inclusion of ‘Dupaningan Agta’ based on one academic work referencing that name, and this has been accepted. [NEXT] And the standard now includes ‘Dupaningan Agta’

[3] http://www.loc.gov/marc/languages/language_code.html (24/1/13)

Further examples of projects which identify languages in some way include Multitree[4], the World Loanword database (WOLD[5]), the World Atlas of Linguistic Structures (WALS), or the Rosetta Project, each with its own ways of dealing with the issue. WALS has geographic information associated with language names and a process for correction of locations in the WALS database.

[NEXT]
Linguasphere is "an exploration into the totality of human languages, observing them as the interactive working parts of a planetary system of communication." It contains 70,900 linguonyms at http://www.linguasphere.info with no readily apparent mapping between the Linguasphere names and other schemes. Of these, 4,500 are 'outer languages'divided into 100 'referential zones' 'within which the modern languages and dialects of the world are classified'. [NEXT]

In particular the notion of a languoid has been developed in Glottolog (http://www.glottolog.org) for which there are 94,049 languoids (Nordhoff 2012: 196) each with their own persistent identifier (Glottolog ID). "Languoid" is a (relatively new) cover term for "language" and "language family". The problem for a search tool is that it is difficult to provide users with prompts for this many languoids, although their website does allow searching (usefully by part [NEXT] or whole [NEXT] of a name, as seen in the screen shots). Further, the languoids are purposely descriptors of various levels of language grouping, from local lect through to language family, and so it will be maximally inclusive, but with the problem that the user has to understand the context in which the term is given (as a higher order group name odown to a dialect name, but this is visually presented in the glottolog search presentation).

[NEXT] WOLD also uses languoids, but the list of items says 'The World Loanword Database contains information on loanwords, source words and other words in 395 languages.'
The help info on language names says 'This is the name of the language (or family, in the case of donor languages) that was adopted in the World Loanword Database. Alternative names can be found on the individual language pages.' So, while the notion of languoid is established, it is not clear how it is implemented.

Good and Hendryx-Parker (2006) in their discussion of the Rosetta Project's approach to what they call 'contested categorization' identify 'lingual nodes' whch could be any one of a language, dialect or family, as they note, "All that is required is for someone to have referred to that entity" (ibid:3).

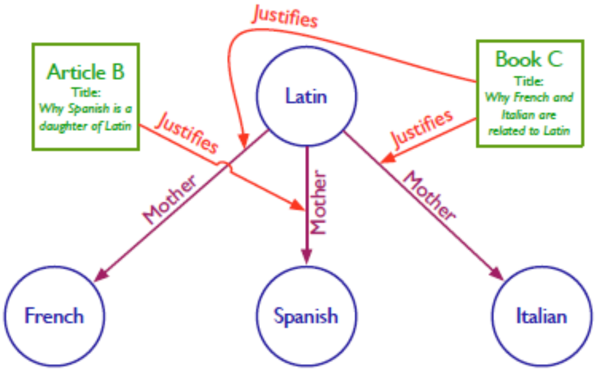
They are concerned with language relationships and genetic groupings (that is, not with the level of language identification except as far as it relates to these groupings) and propose that classifications should be able to be stated in ways that allow different and even contrary versions to co-exist, and for the interpretation of the relationship to change and be mapped over time. A 'node' has three properties:
[NEXT]

[4] http://multitree.org/help (24/1/2013) "Each tree is made up of nodes. A node may be of the type Stock, Subgroup, Language, Dialect or Dialect Group. "

[5] http://wold.livingsources.org/terms#languoid (24/1/2013)

(1) Unique identifier
(2) Metadata (human readable name for the node)
(3) Actual documents in the language (ibid:4)
[NEXT]

The relationships between nodes are mapped using semantic mapping (for which the Resource Description Framework—RDF—is a useful implementation) so that statements like ‘x is a daughter of y’ (called a triple) can be mapped. Further, RDF ‘contexts’ assign metadata to the complete statement or triple and the Rosetta Project assigns a different context class to contested content than it does for non-contested content. A similar process in RDF is ‘reification’ (ibid:15) of a triple that allows it to itself be the argument of a further triple. They illustrate this with the following figure in which the relationship between Spanish and Latin is sourced to Article B



Based on the kind of relationships illustrated above, we could then conceive of an index of the use of standard language codes that included a reliability measure for the purpose to which it is currently being applied. That is, if we have a 1:1 mapping between the language name as identified by the researcher and the standard, then there is either no need for further comment, or the comment could be that the terms are isomorphic, perhaps ‘is the same as’.

Using relationships rather than absolute categories could allow some leeway in the assignment of language identifiers. RELcat has the following relationships:
1. Related (rel:related)
1.1. Same as (rel:sameAs)
1.2. Almost same as (rel:almostSameAs)
1.3. Broader than (rel:broaderThan)
1.3.1. Super class of (rel:superClassOf)
1.3.2. Has part (rel:hasPart)
1.3.2.1. Has direct part (rel:hasDirectPart)
1.4. Narrower than (rel:narrowerThan)
1.4.1. Sub class of (rel:subClassOf)
1.4.2. Part of (rel:partOf)
1.4.2.1. Direct part of (rel:directPartOf)

[NEXT]

In the case of the example AA1-001 from our collection, above, we could state that the term ‘Salako’ is part of the superclass Kendayan (if that is the case). We should probably add a function for users to allow them to select which of the relations holds between the standard term and the one they provide if it is not a perfect match.

So, if the researcher notes that there is no local name in use that resembles the ISO standard then that could perhaps be marked as 'related' but since there may be no known relationship the list given in RELcat may not be sufficient and we may have to include some more terms. Perhaps we could implement a degree of certainty that the code matches the variety – with 1-5 as an attribute of the relationship term where 1 is low certainty that the code or language identified matches the subject language and 5 is a high degree of certainty.

Using the same concept, we could assign relationships to triples so that the relationship of an item in our collection to a language name is assigned a certain degree of reliability. For example, if we suspect some notes are from a given language we could assign a reliability score depending on the source of the information. That is, if the researcher says that the notes are from a given language for which we have a standard code, then we can assign the relationship between their notes and the code to '1. Reliable'. If we have not been able to determine what language the notes are from, but we know the general area the researcher worked in, we could assign '2. Partially reliable'. Finally if we had nothing but the vaguest idea where the notes came from because the researcher was known to have worked on several languages then we could assign '3. Unreliable' to the relationship between the notes and the language name. There would still be value in the assignment of the language name as it will provide a more targetted search than would an absence of a language name. [NEXT]

1. Reliable
2. Partially reliable
3. Unreliable

[NEXT]
Linguasphere[6] uses an asterisk to indicate 'items of data which are unreliable or which require corroboration' (Linguasphere synopsis, p.14)

Conclusion

While it is clearly problematic to attempt mapping the complex nature of people to language to territory relationships and then to assign a standard term to the label for that set, there is still a need for identifying languages. This is especially the case for the kind of resource discovery that is provided by a language archive. Relational terms are offered as a means of providing users with more information about the status of a language name choice in an archival collection.

References

Good, Jeff & Calvin Hendryx-Parker. 2006. Modeling Contested Categorization in Linguistic Databases. In Proceedings of the EMELD 2006 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art. Lansing, Michigan. June 20–22, 2006.

Kamusella, T. 2012. The global regime of language recognition *International Journal of the Sociology of Language* 218:59-86

Nordhoff, S. 2012. Linked Data for linguistic diversity research: Glottolog/Langdoc and ASJP online. pp. 191-200 in Chiarcos, C., S. Nordhoff & S. Hellmann (eds.) 2012. *Linked data in linguistics: Representing and connecting language data and language metadata*. Berlin: Springer.

[6] http://www.linguasphere.info/lcontao/tl_files/pdf/part2/OL-SITE%20Part%202%20Synopsis.pdf