# LINKING RESOURCES, LINKING COMMUNITIES
## An Australian Indigenous Languages database, multimedia projects, and the role of metadata

**Patrick McConvell**
**Australian Institute of Aboriginal and Torres Strait Islander Studies**
For workshop on 'Resources and Tools in Field Linguistics'
at LREC conference, Las Palmas, May 2002

## 1. Introduction

In this paper I describe two projects underway at AIATSIS that have relevance to documentation of Australian indigenous languages.

(a) Development of a national Indigenous Languages database;
(b) The 'Laves Documentation project' which is representative of a class of multimedia projects which we would like to see happen.

In both cases I will describe the projects as wholes but will also focus on the role of metadata in them. Neither of these are fieldwork projects as such but use information from fieldwork and can set the frame for the way field data is collected, including insertion of metadata.

Our perspective is driven by strong participation of Indigenous people in decision-making at the Institute and in other programs with which we work, such as Regional Aboriginal Language Centres, and by Indigenous formulations of models of 'two-way' research ( McConvell 2000, Marika 1999).This means that the aim of language documentation is not solely archiving of data for scientific purposes, but also, and importantly, to make available and adapt such information for language maintenance and educational programs for Indigenous communities. This is in line with the wish of the great majority of Australian Indigenous people that their descendants 'keep their languages going' and revive them if possible.

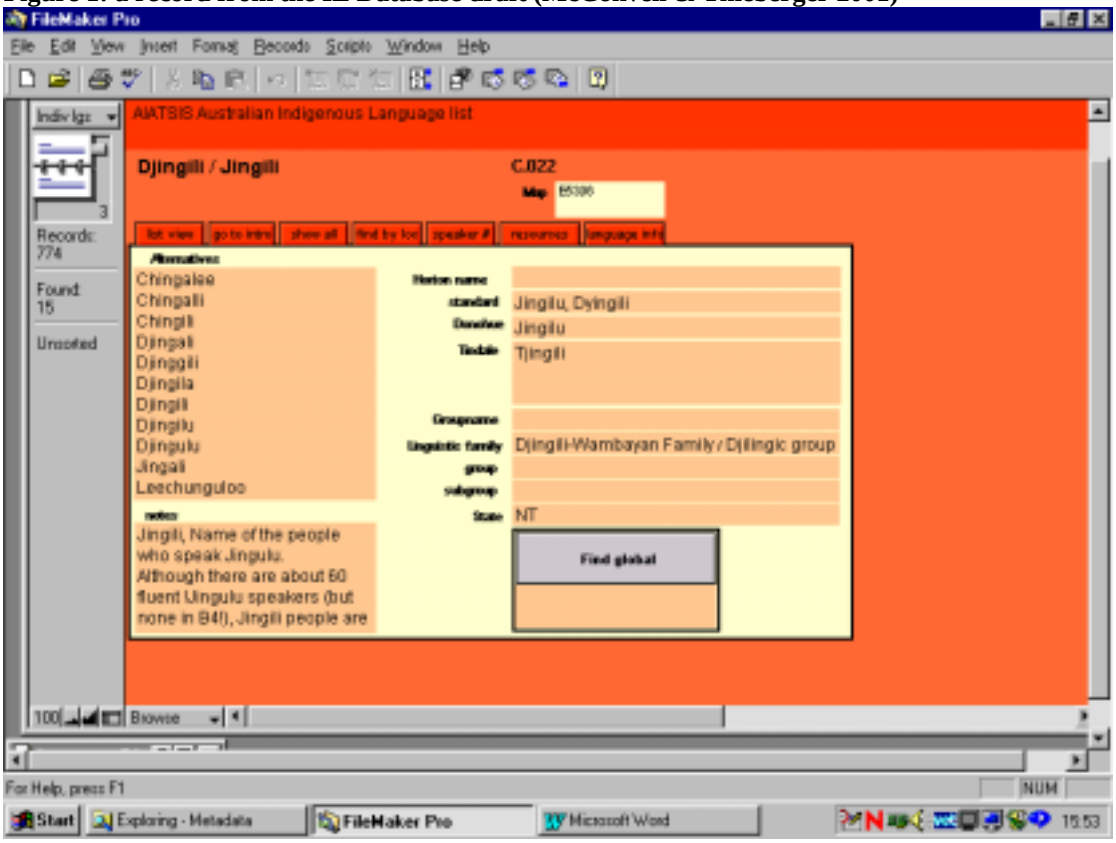
There is a common interest of Indigenous communities and the scientific community that the documentation be as accurate and comprehensive as possible, and compiled and analysed by experts, but the Indigenous community also require that it be accessible and usable by community people with less training and expertise. There may also be a difference in emphasis on content coming from the Indigenous side. While there is recognition that grammar must be analysed, there is more emphasis on the part of Indigenous researchers than the average linguist may give to the cultural context - recording of the detail of old customs and knowledge of the environment, special ways of speaking associated with different kin, ritual relationships and so on.

In the final section of the paper a more general model is presented of how digital resources can be linked together for research - including community-driven research, taking into account this demand from Indigenous co-researchers to link language with culture, land and people.

## 2. A Database of Australian Indigenous Languages

Nick Thieberger compiled a database of Indigenous languages of Australia as part of McConvell and Thieberger (2001; see also summary of findings at 'The state of indigenous languages in cultural heritage' http://www.ea.gov.au/soe/2001/heritage/pubs/part07.pdf ). This is being revised and should soon be more generally available (some of the ways it is being upgraded are touched upon below). It is currently a FileMaker Pro database with the following fields. The 'speaker #'' button leads to a screen giving numbers of speakers according to different standard sources; the 'resources' button gives access to a indicator of the level of documentation according to a scale developed in McConvell & Thieberger (2001) and further discussed below. Languages can be located on a map of Australia, or clicking on a section of the map will yield a range of records.

**Figure 1: a record from the IL Database draft (McConvell & Thieberger 2001)**




The preliminary version of this database is being used by some people in AIATSIS as an authority. Upgrading of the resource is currently being undertaken as outlined in the following sections; implications for general metadata are sketched.

Information for this database is culled from various sources but a series of "Handbooks" of Indigenous languages produced of different regions have been a major inspiration and source of information. These have been produced in conjunction with the work of indigenous-controlled Regional Aboriginal Language Centres, in the main.

## 3. Names and codes for languages

The issue of Australian Indigenous language names is complex. There are copious spelling variations for each form of name in the literature which have to be linked together. These are due to different spelling systems being used, or in many cases, no system. Where practical orthographies have been developed by linguists working with community people, these are the obvious candidates for standard spellings of names but these are far from universally used even by academics.

In addition there are minor or major differences in the form of names of the same group because of free variation or dialect variation in the language, use of foreigner group's name for a language in addition to the group's own name etc. In some cases the preference for an ethnonym or language name among the group itself has changed over the years (see McConvell to appear).

However an alternative authority exists in the AIATSIS Library Thesaurus. A proposal to harmonise these two is being considered and Thieberger is working on this. A Thesaurus gives a list of alternative names/spellings with a standard, but could be configured as a subset of a full listing, in which the relationship of the alternative to the standard term could be more fully described.

On the international scene the *Ethnologue* list of language names and codes has been adopted by OLAC as their authority and source of metadata. Gary Simon has indicated that Ethnologue is willing to adapt its listing based on recognised country standards and regularly upgrade in this fashion also (Simons 2002), and we hope to reach a point of direct translation/compatibility between the AIATSIS listing and codes and *Ethnologue* shortly.

A secondary problem arises because of the fact that *Ethnologue* lists mainly only 'living' languages and some 'recently extinct' languages, where 'recently' is not further defined, selected on the basis of whether they are considered in some respect significant by linguists or have Scripture published (Grimes ed. 2000:viii) A number of languages of interest both to linguists and to Indigenous communities in Australia (and I would guess elsewhere) are not listed because they are 'extinct'. They can still be worked on from previous records and remembered fragments in some cases, and a number are actively being revived. A case in point is Kaurna of the Adelaide region (Amery 2001): absent from the Ethnologue, it has been added to the complementary list of extinct languages being compiled by E-MELD, but it is hoped that this listing will also make an effort to keep in harmony with recognised national standards such as the IL database. Another related issue is that descendants of the Kaurna and similar groups reject the terms 'extinct' and 'dead' for their languages and call them 'sleeping', because in their view they can be revived. While the term 'sleeping' is probably too local for general international use, a more neutral term for languages not currently spoken would be valuable: I suggest 'not spoken'.

## 4. Metadata for language names

Metadata schemes of most relevance to governmental regulation of the functions of AIATSIS are AGLS (Australian Government Locator Scheme) and NLA (National

Library of Australia. OLAC is the most widely promoted scheme for languages, certainly for small and indigenous languages. The following chart compares the schemes (prepared by Mark Denbow of Audio-Visual Archives, AIATSIS).

For both language and subject.language OLAC supplied the *Ethnologue* list as its controlled vocabulary, hopefully to be adapted to include IL database refinements as discussed above.

**Figure 2:Comparison of Metadata ElementsWithin Different Schemes**

| *AGLS* | *AGLS Status* | *Dublin Core* | *OLAC* | *Thesauri or Encoding Scheme* | *Optional* |
|---|---|---|---|---|---|
| Creator | Mandatory | Creator | Creator | Commonwealth Govt On-Line Directory (GOLD) | |
| Date | Mandatory | Date | Date | ISO 8601 | |
| Description | Mandatory | Description | Description | | |
| Title | Mandatory | Title | Title | | |
| Type | Mandatory | Type | Type | AGLS Document<br>AGLS Service | |
| Subject | Choose One | Subject | Subject | Aust Public Affairs Information Service (APAIS) | LCSH |
| Identifier | Choose One | Identifier | Identifier | URI<br>DOI<br>ISBN<br>ISSN<br>USID | OAI |
| Publisher | Conditional | Publisher | Publisher | GOLD | |
| Coverage | Conditional | Coverage | Coverage | Date – ISO 8601<br>Location - LCSH | TGN |
| Language | Conditional | Language | Language | RFC3066<br>ISO639 Lang +<br>ISO3166 Country | |
| Contributor | Optional | Contributor | Contributor | GOLD | |
| Relation | Optional | Relation | Relation | | |
| Rights | Optional | Rights | Rights | | |
| Source | Optional | Source | Source | URI/URL<br>ISBN / ISSN<br>USID | |
| Format | Optional | Format | Format | IMT + IANA | |
| Mandate | Optional | | | | |
| Audience | Conditional | | | | |
| Availability | Choose One | | | | |
| Function | Choose One | | | | |
| | | | | | |
| | | | | | |
| | | | | | |
| | | | Subject.language | | |
| | | | Format.cpu | | |
| | | | Format.encoding | | |
| | | | Format.markup | | |
| | | | Format.os | | |
| | | | Format.sourcecode | | |
| | | | Type.functionality | | |
| | | | Type.linguistic | | |
| | | | | | |

## 5. Levels of endangerment of Australian Indigenous languages

Many suggestions have been made about how endangerment should be measured; the most useful are those which regard an endangered language as one which is not spoken by children of the group, or only by a small number of the children (Kinkade 1997 ; Wurm 1996). Further divisions can be recognised on the basis of whether other older age-groups also do not speak the language. One such scheme has been proposed by McConvell & Thieberger (2001) and taken up by McConvell et al. (2002), as shown in the chart below:

**Figure 3: Levels of endangerment**

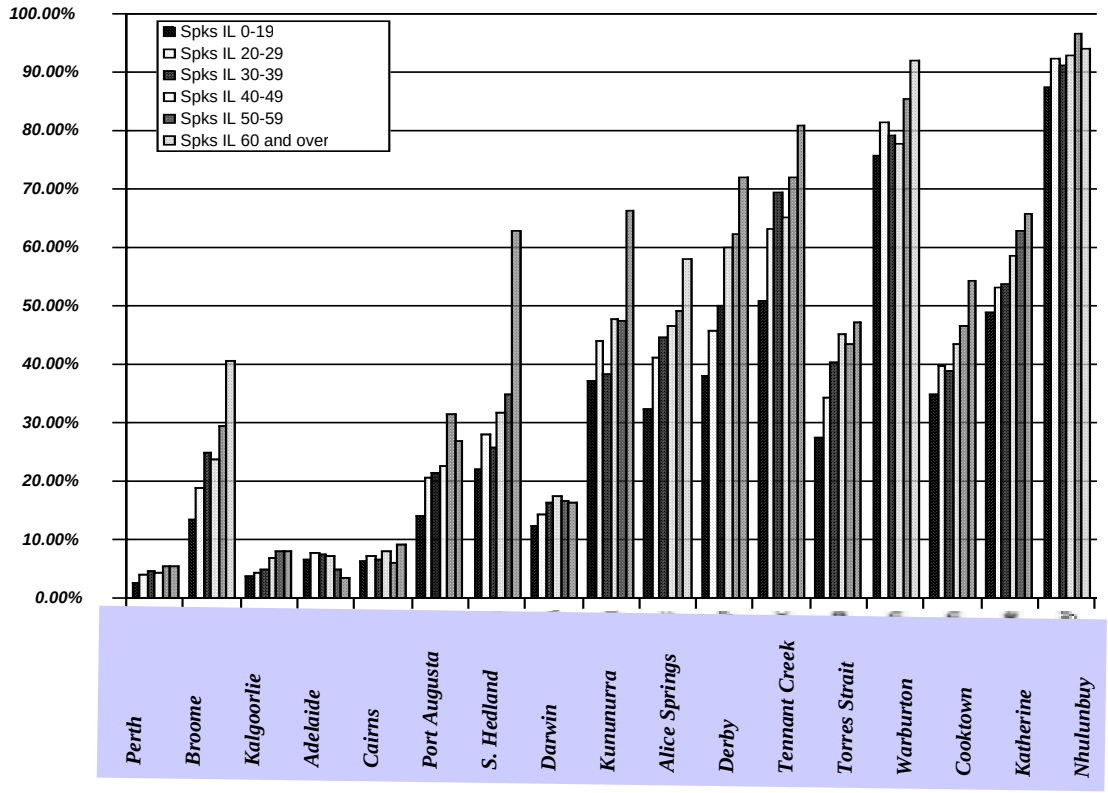
| Age | Strong | Endangered (Early Stage) | Seriously Endangered | Critically endangered ('Near-Extinct') | Not Spoken ('Extinct') |
|---|---|---|---|---|---|
| 5-19 | speak | don't speak | don't speak | don't speak | don't speak |
| 20-39 | speak | speak | don't speak | don't speak | don't speak |
| 40-59 | speak | speak | speak | don't speak | don't speak |
| 60+ | speak | speak | speak | speak | don't speak |

The terminology using 'extinct' or 'dead' is not favoured by Australian Indigenous groups and has been replaced by other terms here: 'critically endangered' for 'near-extinct' and 'not spoken' for 'extinct' . 'Speak' is construed as meaning 'can understand and produce coherent sentences with appropriate vocabulary and grammar approximating to that of older people on a range of topics'. Following Wurm and Drapeau, 'don't speak' is interpreted as meaning that less than 30% of the population regarded as affiliated to the language have that ability, except in the case of 'not spoken' where 'don't speak' is interpreted as a situation where noone speaks the language ('speaking' being defined as above).

Analysis of census data from 1996 (of which Figure 4 is a part) together with the results of a survey conducted by ATSIC in 1994 shows that patterns of language use across age groups fall into five main patterns of which one probably results from interference between different language groups with different degrees of endangerment. Otherwise the main patterns are the first three shown in Figure 5, which correlate roughly with the designations 'strong'; 'endangered' (early stage or serious); and 'critically endangered' or 'not spoken'. The anomalous pattern is that of Adelaide which probably relates to language revival in that city and the surrounding region.

However because of the way that census data is collected in Australia it is not possible to derive endangerment indices for individual languages from that data alone. It is however possible to make assessments from other data and include approximate indices for individual languages in the IL database.

**Figure 4: Age groups speaking Indigenous languages by region (sample only)**



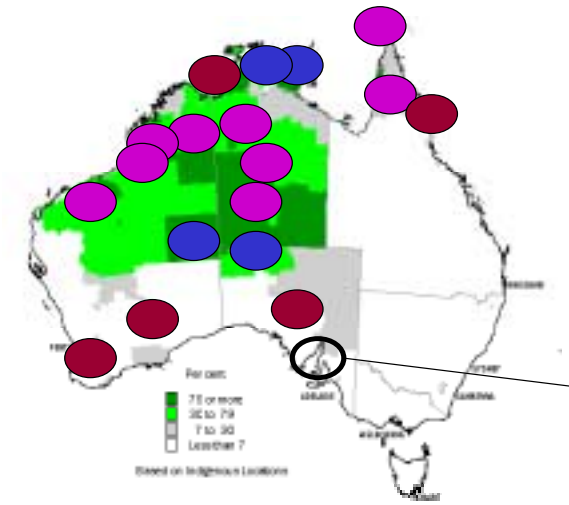

## *AGE GROUPS SPEAKING IL's (by region)*

(ABS 1996 Census)

.

.

**Figure 5:Patterns of Language Shift**



***PATTERNS OF LANGUAGE SHIFT***

***•Nearly all speak IL at home; not much difference between age groups.***

***•Many less speak IL at home than claim to know it ;sharply declining use of IL among young.***

***•Very few speak IL at home; proportion of speakers low in all age groups and declining.***

**•Adelaide: None speak IL at home (1994); IL knowledge increasing among young (1996 census).**

## 6. Levels of documentation

In the IL database each record which pertains to a language is linked to a record of the documentary resources about the language.

To achieve measurable indicators a points system was built into this system. In McConvell & Thieberger (2001) we have implemented a point system to describe the documentation of a languages as follows (with a possible total of 17 points for a well-documented language):

**Figure 6:Indicators of documentation**

**Dictionaries**: Detailed dictionary (e.g. Arrernte, Kayardild) (4); Medium dictionary (3); Small dictionary/ wordlist (e.g. Warnman) (2); Simple wordlist (e.g. Bates, Curr) (1).

**Texts**: Extensive text collection (3); Several texts (<10) (2); Elicited/example sentences (1).

**Grammar**: Detailed grammar (e.g. Gooniyandi, Kayardild) (4); Middle-sized grammar (eg. Handbook) (3); Grammar sketch or many technical articles (2); Few technical articles only (1).

**Ethnolinguistic information**: Substantial ethnolinguistic work (e.g. thesis) (3); Ethnolinguistic description (2); Some ethnolinguistic information (1).

**Audio recording**: More than several hours of audio (3); Less than several hours of audio (2); Less than an hour of audio (1); No audio recorded (0).

However there are some other resources in or on languages which are also of importance on language which we recommend be included, to produce a 20-point system:

**Other:** Literature (including school readers and religious translation) in the language - more than 1000 words (2); more than 100 words (1); video or film with more than 100 words spoken or subtitled **or** multimedia with more than 100 words spoken and/or written (1).

It is obviously beneficial if metadata categories especially *type.linguistic* harmonised with the categories used for measurement here. In that way a search on a comprehensive database of languages would automatically yield documentation indices.

In the recent South Australian language needs survey (McConvell et al. 2002) a formula involving level of endangerment and level of documentation is proposed to give a rough guide to priority for work - high endangerment and low documentation yields high priority.

## 7. Language rights and protocols

A further important type of metadata is *rights* , concerning rights to and access to data. This is mainly conceived in terms of copyright and the laws of the governments converned with this. However for Indigenous people, and increasingly impinging on national and international legal systems, is the question of Indigenous Intellectual Property rights arising from their own laws and customs. In Australia, Native Title has been recognised in Australian law since 1992 and arguably this also has implications that language groups have common law rights to exercise their own protocols over some uses of their language and their own intellectual products in the languages. This area is far from clear even among Indigenous advocates of this position, but it is being worked out at present.

At the moment *Rights* is treated as a fairly free field in which information on contributors to a piece of data may be listed and may perhaps include restrictions on access, notes on inheritance etc. However it is clear that while idiosyncratic information should be noted in the metadata for a resource, there are in fact existent or emerging principles involved here which are broader. At least some of these principles may be located at the 'language group' level and may therefore be appropriately be placed in such a location as the IL database, to be called up by resources that belong to that language.

**8. Database of IL Programs**

There is also a second database of Indigenous Language programs (educational and community) which is linked to the above database. This needs more work to make it functional.

**9. Linking different types of data**

The above sections are mainly concerned with how a database of Indigenous languages can provide an authority for metadata for data resources on languages, and can call up information about the language, including its endangerment and documentation status, and possible language rights and protocols, when records of the resource are accessed. In this section I discuss further how different types of data might be linked together at least in part using metadata to do so.

AIATSIS is beginning digitization of print and audio-visual materials. It has been decided that the digital products of this will not be part of ASEDA, and in future only language related materials will be held in ASEDA. The other sections of AIATSIS which are bigger players in digitisation are the Library and Audio-visual Archives. Up to recently Library dealt with print materials and Archives with sound and vision (largely analog tapes) and the products of digitization will go back to the section associated with their original medium.

More and more of the research materials we receive are ‘born-digital’ however, and while we might try to assign them a category based on whether they can be produced as a printed document or a sound or picture, this is increasingly artificial. Many databases cannot be printed at all, and multimedia products typically have attributes of all the categories together.

A radical approach would be to abolish all these heritage distinctions and just operate with a large undifferentiated digital library/archive as many institutions do – as ASEDA was originally designed. Differentiation into audio, text (and other potentially useful kinds) is written into the metadata on items and can be used when necessary. We have to live in the real world however where this cannot be achieved, certainly not perhaps for years to come.

What to do in the meantime? Obviously there is a benefit from having available an IL database of the type we are developing and maximising links between that and any resource items in the catalogue, as discussed above.

Beyond that though, one important thing about the kind of data held in the AIATSIS library, archive and ASEDA is that there are enormous numbers of links between the items in different locations, few of which are currently retrievable. For instance, there could be an audio-tape of a text, a digital text version of a transcript, a wordlist or dictionary of the language, digital or print, photographs of the speaker, maps of the locations referred to, genealogies of the speaker or referents etc. with no obvious way of finding one from the other.

Clearly metadata can provide much of the answer to this kind of problem. Within the proposed schemes referred to above (including OLAC) only 'Relation' seems to provide scope for providing this kind of linking metadata. Nobody however is going to approach the whole task of providing such linking metadata all at once: it is vast undertaking. With our limited resources we have to show the value of providing rich linking metadata by doing projects that show the value of doing this. Another project which is going on in Australia which has been better funded and is carrying on parallel work is devoted to making the kinds of links that Indigenous communities want – the Ara Irititja project working with the Pitjantjatjatjara/Yankunytjatjara people of northern South Australia and neighbouring regions.

**10. Laves Digitization Project**

Within AIATSIS we are beginning one project which is in a small way going to progress some of these lines of thinking as well as hopefully produce a very valuable product for researchers, including Indigenous community researchers. This will focus on the work of Gerhardt Laves. The initial stages will involve production of digital graphic images of his fieldnotes and keyboarding and interpretation of the notes as digital text. These two sets need to be linked through metadata. Beyond that there could be links to the few musical recordings he made, genealogies, maps and other images. All this will be carried out under the eye of appropriate indigenous advisers, providing also input into the 'Rights' fields which might control access. Since the data also provides a picture of several different dialects of a language no longer spoken in traditional form (Nyungar) it could also be built into a project that provides a multidialectal dictionary project. Working out how to organise the metadata links within and beyond this project will be the first step in an ongoing process.

Gerhardt Laves was a brilliant US linguist/ethnographer who travelled extensively in Australia in 1929-31 gathering large amounts of data on Indigenous languages and cultures in a number of areas. The materials are mainly handwritten fieldnotes. They are of excellent quality and are extremely valuable for a number of purposes including endangered and in some cases extinct language documentation; native title and family history (genealogies and stories are included). They were overlooked for decades in the US and only in the 1990's have they been gathered together at AIATSIS and begun to be appreciated, in part through the efforts of David Nash, who is currently working part-time managing ASEDA and the Laves project. They are difficult to access and read in their current form owing to factors including faintness of the writing and special symbols used by Laves. They are in urgent need not only of digitization into graphic images, but also keyboarding so that they can be republished in a more accessible text format which is searchable.

The Research section is beginning a ‘Laves Digitization Project’ in collaboration with the Library to create a digital text version of Laves’ work, suitable for print publication and/or web or CD-ROM, including other elements (maps, genealogies, images, sound, as appropriate).

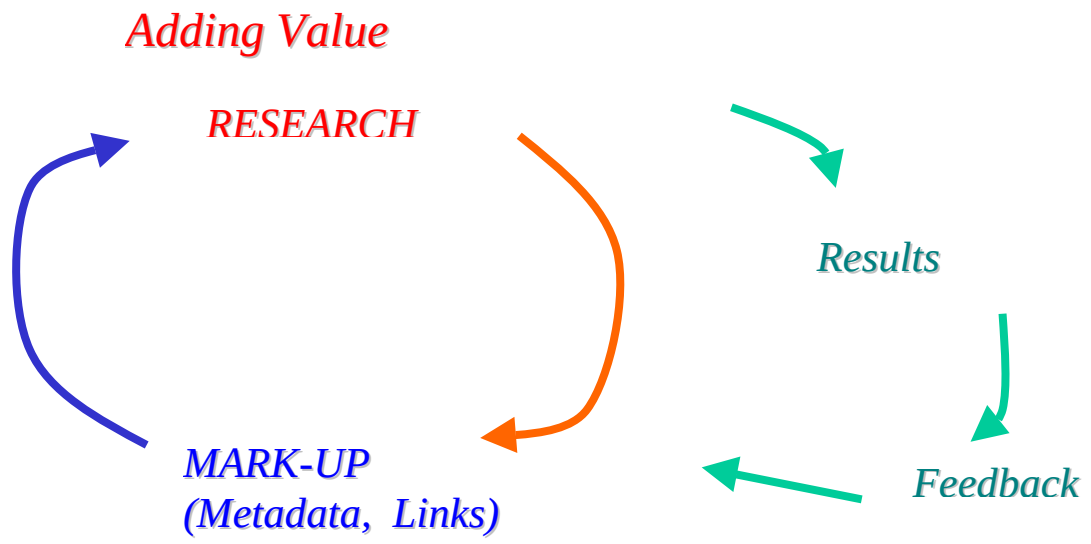
## 11. Ideas for the future

There is potential to build on this and similar projects, bearing in mind some basic principles (cf. McConvell 2000,2001) e.g.

> Knowledge exchange is a social relationship
> Don’t let technology drive you; you drive technology
> Digital data can provide powerful tools for research
> Indigenous community people also do research

A simple model in mind is the following (Figure 7). Research is carried out and the results are marked up as far as possible with metadata which will make them accessible, at which point they may reenter the research cycle. At the same time the results are subject to feedback both from the academic and Indigenous communities - a process which itself adds to and clarifies what metadata and links are needed to pursue further research which may be of the classic academic linguistic type or may be for cultural maintenance purposes.

**Fig.7 Adding value to research products**




The multimedia phase of the Laves project would involve linking of elements within the transcribed texts of languages with other elements deriving from either Laves's work, other archives, or additional information collected for the purpose. This is shown graphically (and loosely) on Figure 8. The types of data contained in this network of data include

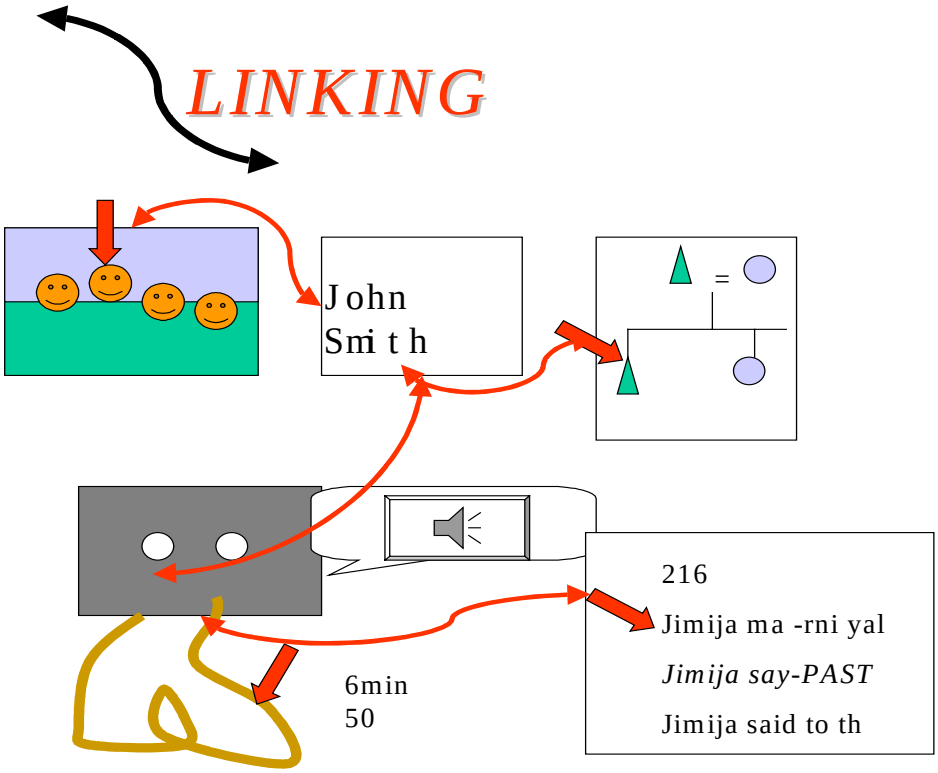
interlinear text with English, morphological break-up and an Indigenous language;
name of a person in a list (database) of people including specification of equivalence between an Aboriginal name and an English name
a photograph (if available) in which one object (person) within the photograph is linked to a name in a list (as implemented in the *Ara Irititja* project)
a genealogy such as occurs in Laves' fieldnotes enhanced by other information. This will probably require specialised genealogy software.

Links to a dictionary, maps and other types of databases would also come into play and could be built on in phases.

The idea here would be that people could enter the network in a number of different ways which would be suitable for different users eg via a name or photograph, or via text in a language or in English, so that both community users and linguists, even dilettante searchers for specific features, could be accommodated.

How exactly such a network would be constructed is something to be worked out in detail, and we are keen to see how other documenters of languages are coping with such problems. Obviously this is in a sense a secondary product based on linguistic and other more basic documentation, but being aware of this kind of product being a likely outcome will affect what kind of metadata fieldworkers might add .

**Fig 8: Linking data in a multimedia project**




Below, in Figure 9, I give a second example of how a multi-media linked network could be built up, this time based on developing interdisiplinary work being done on Aboriginal artifacts (McConvell and Smith to appear; Akerman and McConvell 2002) another endangered field of knowledge which is also considered important by Indigenous people.

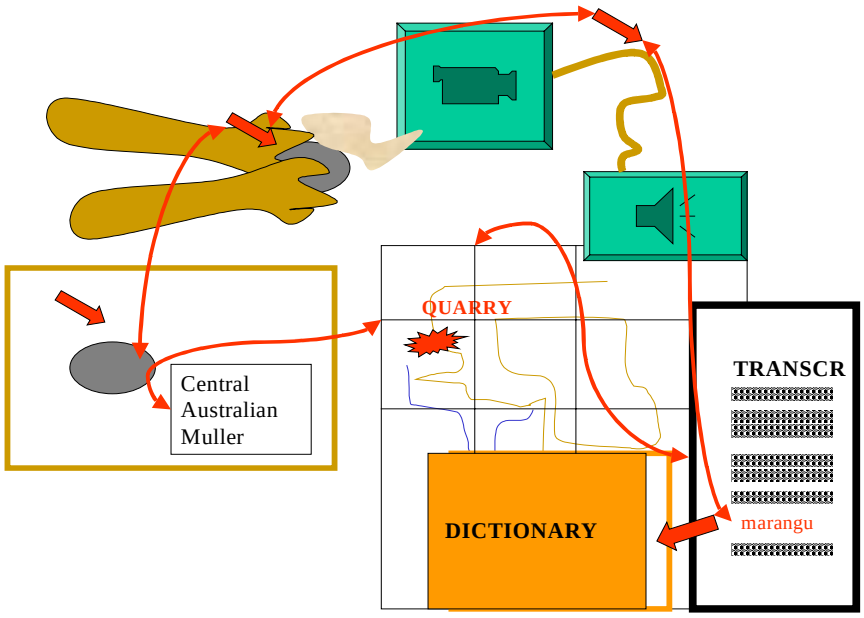
The elements shown on the diagram are

the artifact (in a list of artifacts) and /or still photographs/drawn diagrams
functional descriptions (in this case an upper grindstone or 'muller' is used as an example)
map/location specifications of the quarry from which the stone came
video footage of use of the artifact and/or its manufacture
the sound track of the above or other audio description
transcription of the above
a dictionary of the relevant language with a link between terms for tools and processes in the text and dictionary entries

Additional elements which could be added could be locators for the artifact in Museum catalogues (via URL's perhaps) and links to other databases which have been constructed (stone or timber databases for instance).

Once again this allows multiple entry points for different types of researchers including Indigenous people.

**Figure 9:Linking data in a multimedia network on artifacts**



## 12. Conclusions

This paper has described some projects which are underway at AIATSIS which aim to deliver enhanced value to research products for Indigenous community researchers and academics, and to encourage the co-participation of such groups in teams in the research enterprise. In pursuing this aim digital technology has the potential, if handled correctly, of breaking down the barriers which have prevented Indigenous people from using the results of research and building bridges which can only help all researchers in the long run. It is not the aim to serve up only a fraction of the riches of the resources available, based on some notion (often mistaken) of the kinds of things that Indigenous people are interested in. Rather these differing priorities can be used to design ways into digital resource networks which allow for multiple paths to be fnavigated. In order to make this possible it is necessary to mark up research products more carefully and with broader uses in mind than those which have often dominated linguistics. In some cases, the laboriousness of this task can be reduced by having standard authority databases which can supply the relevant metadata, such as the Australian IL database discussed here. In other cases a useful approach would be to designing multi-media interdisciplinary products specifically with the idea of encountering and finding solutions to the problems involved. Our ability to do this at AIATSIS is constrained by lack of funds for this kind of project, which is not generally seen as part of the research task. This is where partnerships with other bodies in Australia and overseas, and sharing solutions outside the realm of commercial software development is crucial.

## References


Akerman, K and P.McConvell (2002) *Wommera:* the technology and terminology of spearthrowers in Australia. Paper to Centre for Research on Language Change seminar, Australian National University.

Amery, R. (2001) *Warrabarna Kaurna! Reclaiming an Australian language.* Lisse: Swets & Zeitlinger.

Drapeau, L. (1998) Aboriginal languages: current status. In Edwards J ed. *Language in Canada*. Cambridge University Press. 144-159.

Grimes, B. ed.(2000) *Ethnologue; Volume 1 , Languages of the World.* 14th edition. Dallas Texas: SIL International.

Kinkade, M.Dale (1991) The decline of native languages in Canada. In Robins & Uhlenbeck eds *Endangered languages.*157-176. Oxford: Berg.

Marika, R. (1999) Milthun Latju Waanga Romgu Yolngu: Valuing Yolngu Knowledge in the education system. *Ngoonjook.* 16: 107-120.

McConvell, P. (2000) Two-way research resources for Indigenous Languages: positioning resources in the *Garma.* Paper to Linguistic Exploration, Workshop on

web-based Language Documentation and Description.
www.ldc.upenn.edu/exploration/expl2000

McConvell, P. (2001) Looking for the two-way street: Indigenous Australians battle to keep their languages strong.*Cultural Survival Quarterly* 25.2: 18-22.

McConvell, P (to appear) Linguistic Stratigraphy and Native Title: the case of Ethnonyms. In J.Henderson and D.Nash eds. *Linguistics and Native Title.* Canberra: Aboriginal Studies Press.

McConvell, P. and M.Smith (to appear) Millers and mullers: the archaeolinguistic stratigraphy of seed-grinding in Central Australia. In H.Andersen ed. *Linguistic stratigraphy and prehistory.* Amsterdam: John Benjamins.

McConvell, P and N.Thieberger (2001) The State of Indigenous Languages in Australia. Environment Australia Technical Paper, not yet published.

McConvell, P., R.Amery, M-A Gale, C. Nicholls, J. Nicholls, L.I.Rigney, S.U.Tur (2002) *"Keep that language going!" A needs-based review of the status of Indigenous Languages in South Australia.* Canberra: AIATSIS.

Simons, G. (2002) SIL 3-letter codes for identifying languages. LREC International Workshop on Resources and Tools in Field Linguistics.

Wurm, S A (1996) *Atlas of the world's languages in danger of disappearing.* Paris/Canberra: UNESCO Publishing/Pacific Linguistics