LANGUAGE IDENTIFYING CODES: REMAINING ISSUES, FUTURE PROSPECTS

Simon Musgrave (Monash University, Australian National Corpus)
Linda Barwick (PARADISEC, University of Sydney)
Michael Walsh (AIATSIS, University of Sydney)
Andrew Treloar (Australian National Data Service)

Overview

- The importance of accurately identifying languages
  - Within research communities
  - In a wider context
- ISO639-3 – working with what we have
  - Improving linguistic input – Australian examples
  - Improving processes – registration authorities
- Looking to the future
  - Current developments in ISO639/TC37
  - Ways of influencing outcomes:
    - ARGILaRe
    - RDA Working Group

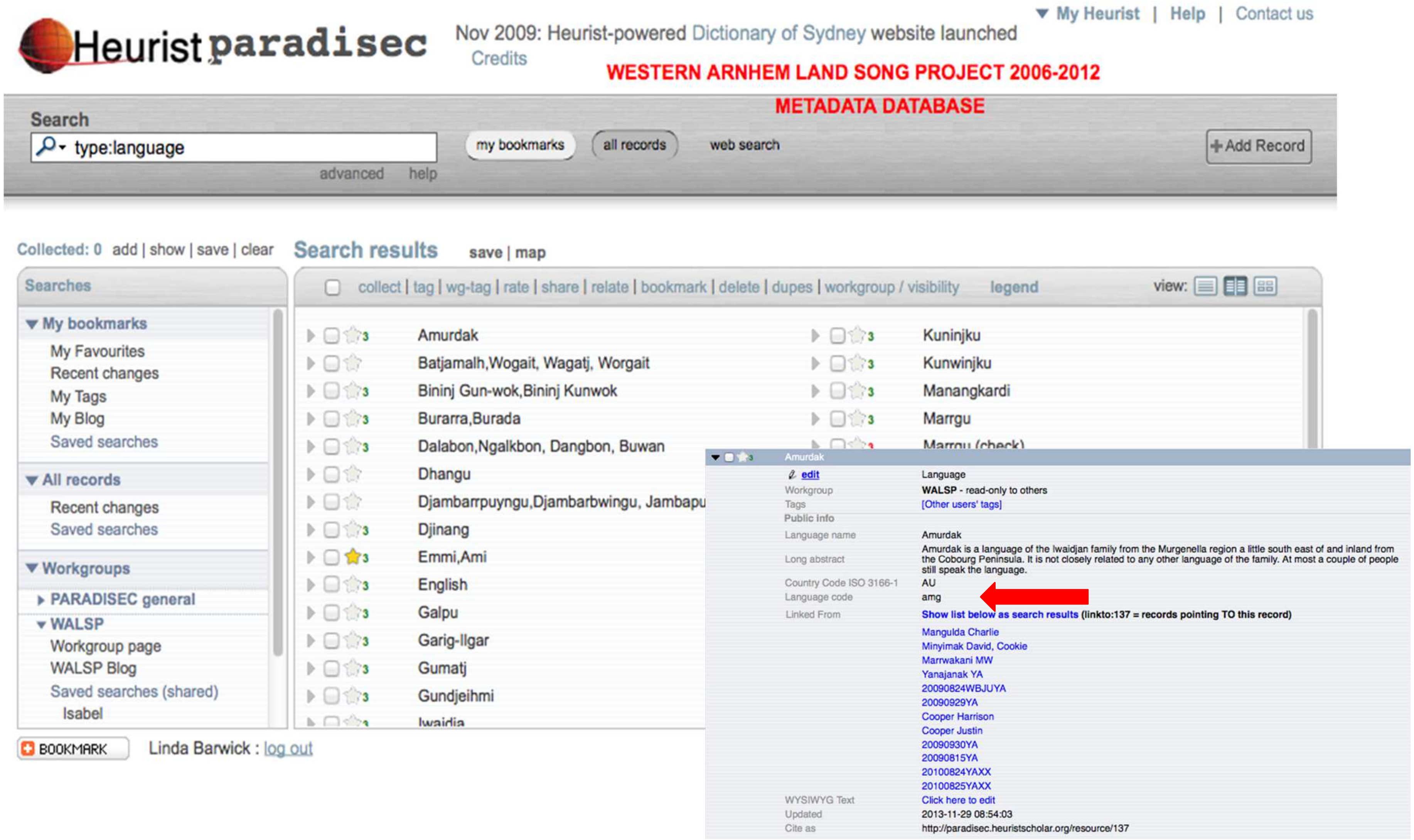
Language codes in disciplines outside linguistics

- Language coding could improve discovery and accessibility of web and library resources in other disciplinary areas e.g.:
  - Song
  - oral history
  - (ethno)biology
  - ...
- Accurate identification of languages can facilitate reintegration of cultural knowledge across domains
- This is a benefit for public and researchers

Implementation example

- Particularly useful for organizing and management of multilingual collections
- E.g. Western Arnhem Land song project
  - ISO639-3 (language identifiers) already adopted
  - potential use for 639-5 (language families and groups)
  - decisions on 639-6 (language variants) may be impossible without linguistic advice
- Only if appropriate infrastructure available:
  - standards agreed
  - tools for coding available
  - platforms for aggregation created, etc
- Need for dialogue with the library/cataloguing community

42 Australian languages in Western Arnhem Land Song corpus



Amurdak
Batjamalh,Wogait, Wagatj, Worgait
Bininj Gun-wok,Bininj Kunwok
Burarra,Burada
Dalabon,Ngalkbon, Dangbon, Buwan
Dhangu
Djambarrpuyngu,Djambarbwingu, Jambapu
Djinang
Emmi,Ami
English
Galpu
Garig-Ilgar
Gumatj
Gundjeihmi
Iwaidja

Kuninjku
Kunwinjku
Manangkardi
Marrgu
Marrgu (check)

Amurdak

edit	Language
Workgroup	**WALSP** - read-only to others
Tags	[Other users' tags]
Public info	
Language name	Amurdak
Long abstract	Amurdak is a language of the Iwaidjan family from the Murgenella region a little south east of and inland from the Cobourg Peninsula. It is not closely related to any other language of the family. At most a couple of people still speak the language.
Country Code ISO 3166-1	AU
Language code	amg
Linked From	**Show list below as search results (linkto:137 = records pointing TO this record)** Mangulda Charlie Minyimak David, Cookie Marrwakani MW Yanajanak YA 20090824WBJUYA 20090929YA Cooper Harrison Cooper Justin 20090930YA 20090815YA 20100824YAXX 20100825YAXX
WYSIWYG Text	Click here to edit
Updated	2013-11-29 08:54:03
Cite as	http://paradisec.heuristscholar.org/resource/137

Language coding for song

- Special song languages/registers: coding 'spirit language' used in song; coding songs entirely in vocables
  - Could use ISO639-3 mis ‘uncoded languages’
- Granularity: code switching within a single song item (e.g. 4 languages in single Malgarrin text

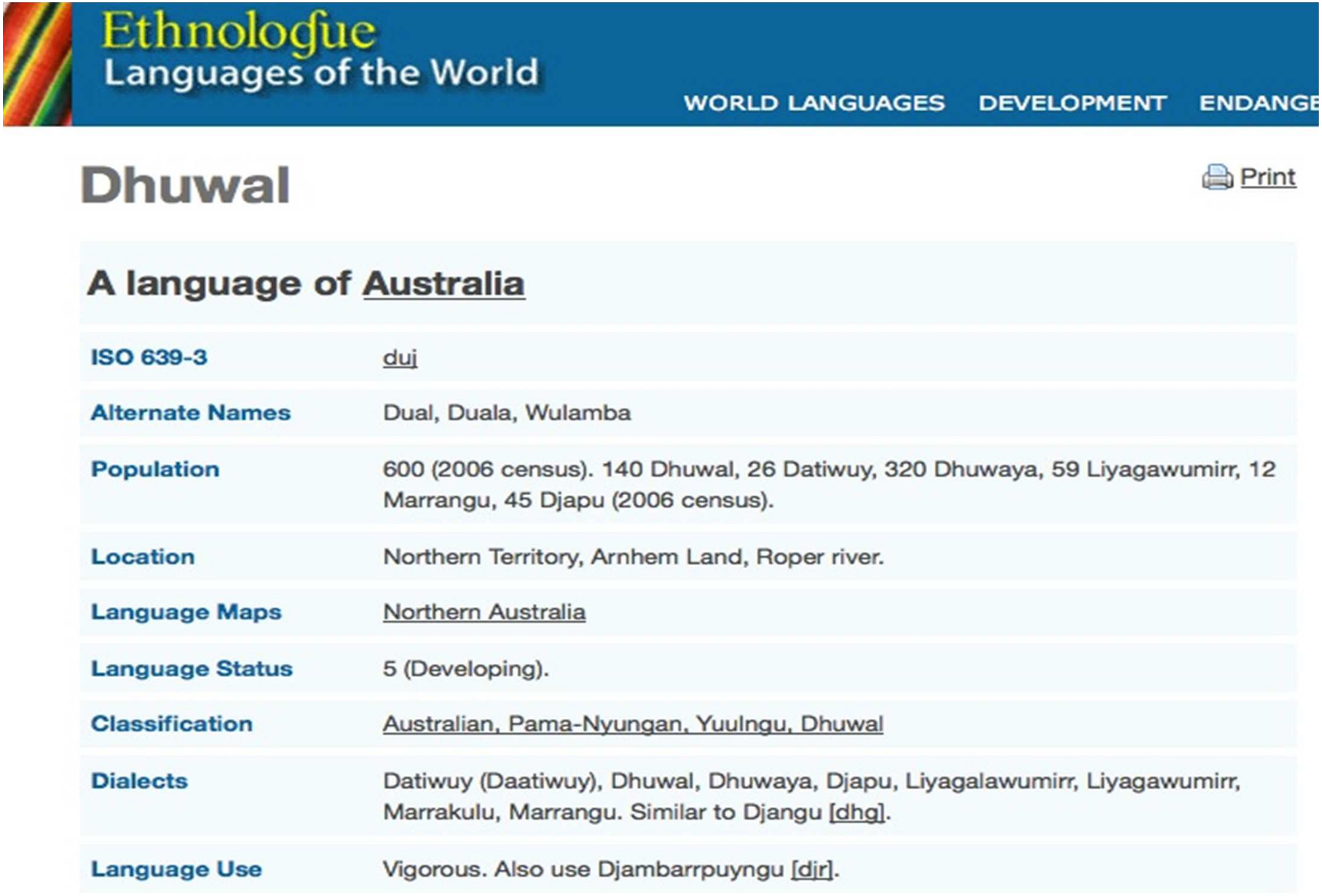
mulurn	**kanarra**	**puratj**	**parraya**	**wantinya**
shade/leaves	*leaves*	*brush*	*you go*	*[unknown]*
Murrinh-patha mwf	Gija (Kitja) gia	English eng	Gija gia	Djaru (Jaru) ddj

- Indeterminacy: language-distinguishing grammatical markers may be absent, though possible to determine higher order language grouping (Arandic, Bininj Kunwok)

Wider communities

- Mac Developer Library:
  For language designations, you can use either the ISO 639-1 or ISO 639-2 conventions.
  (https://developer.apple.com/library/mac/documentation/macosx/conceptual/bpinternational/Articles/LanguageDesignations.html)
- Osborn, D. 2010. *African Languages in a Digital Age*. Cape Town: HSRC Press.
  This set of standards [ISO639] serves several purposes, including the identification of the languages of web content and the selection of appropriate locale information. (p73)
- W3C Internationalization. 2009. **Language tags in HTML and XML**
  All language tags must begin with a primary language subtag……These codes come from, and are kept up to date with, ISO 639 language codes.
  (http://www.w3.org/International/articles/language-tags/)

Yolngu: http://www.ethnologue.com/language/duj

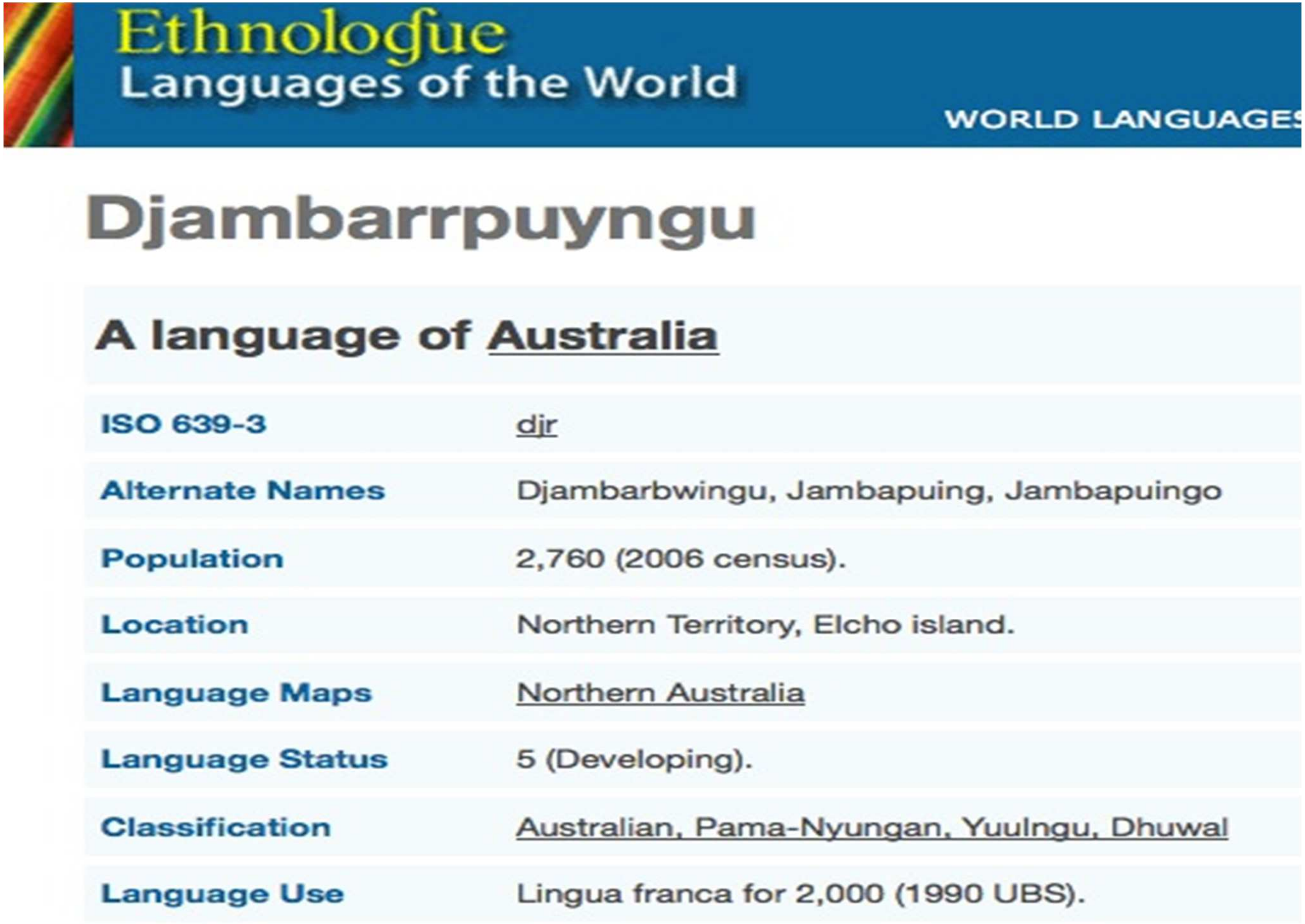


Dhuwal

A language of Australia

ISO 639-3	duj
Alternate Names	Dual, Duala, Wulamba
Population	600 (2006 census). 140 Dhuwal, 26 Datiwuy, 320 Dhuwaya, 59 Liyagawumirr, 12 Marrangu, 45 Djapu (2006 census).
Location	Northern Territory, Arnhem Land, Roper river.
Language Maps	Northern Australia
Language Status	5 (Developing).
Classification	Australian, Pama-Nyungan, Yuulngu, Dhuwal
Dialects	Datiwuy (Daatiwuy), Dhuwal, Dhuwaya, Djapu, Liyagalawumirr, Liyagawumirr, Marrakulu, Marrangu. Similar to Djangu [dhg].
Language Use	Vigorous. Also use Djambarrpuyngu [djr].

http://www.ethnologue.com/language/djr



Ethnologue
Languages of the World

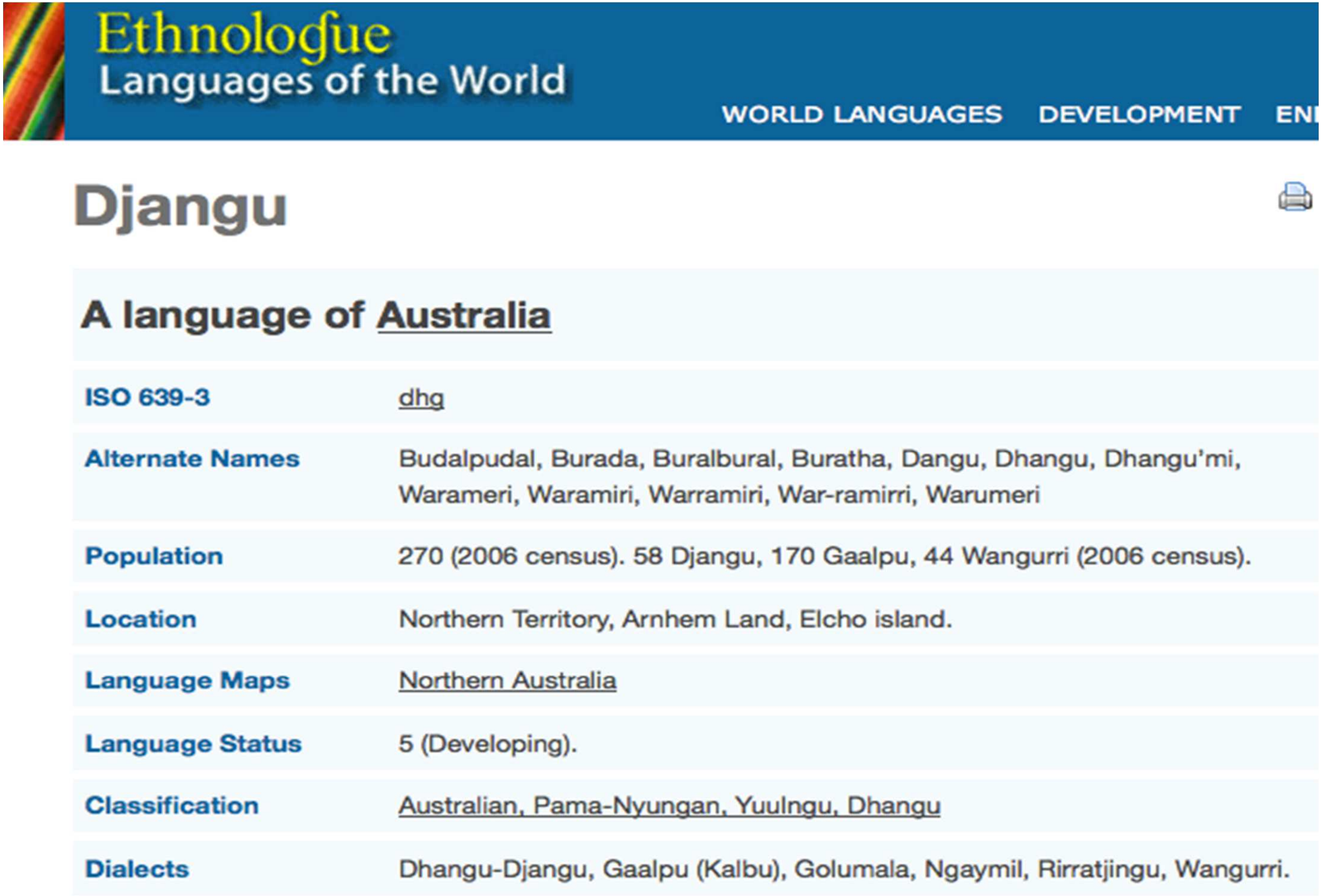
WORLD LANGUAGES

Djambarrpuyngu

A language of Australia

ISO 639-3	djr
Alternate Names	Djambarbwingu, Jambapuing, Jambapuingo
Population	2,760 (2006 census).
Location	Northern Territory, Elcho island.
Language Maps	Northern Australia
Language Status	5 (Developing).
Classification	Australian, Pama-Nyungan, Yuulngu, Dhuwal
Language Use	Lingua franca for 2,000 (1990 UBS).

http://www.ethnologue.com/language/dhg

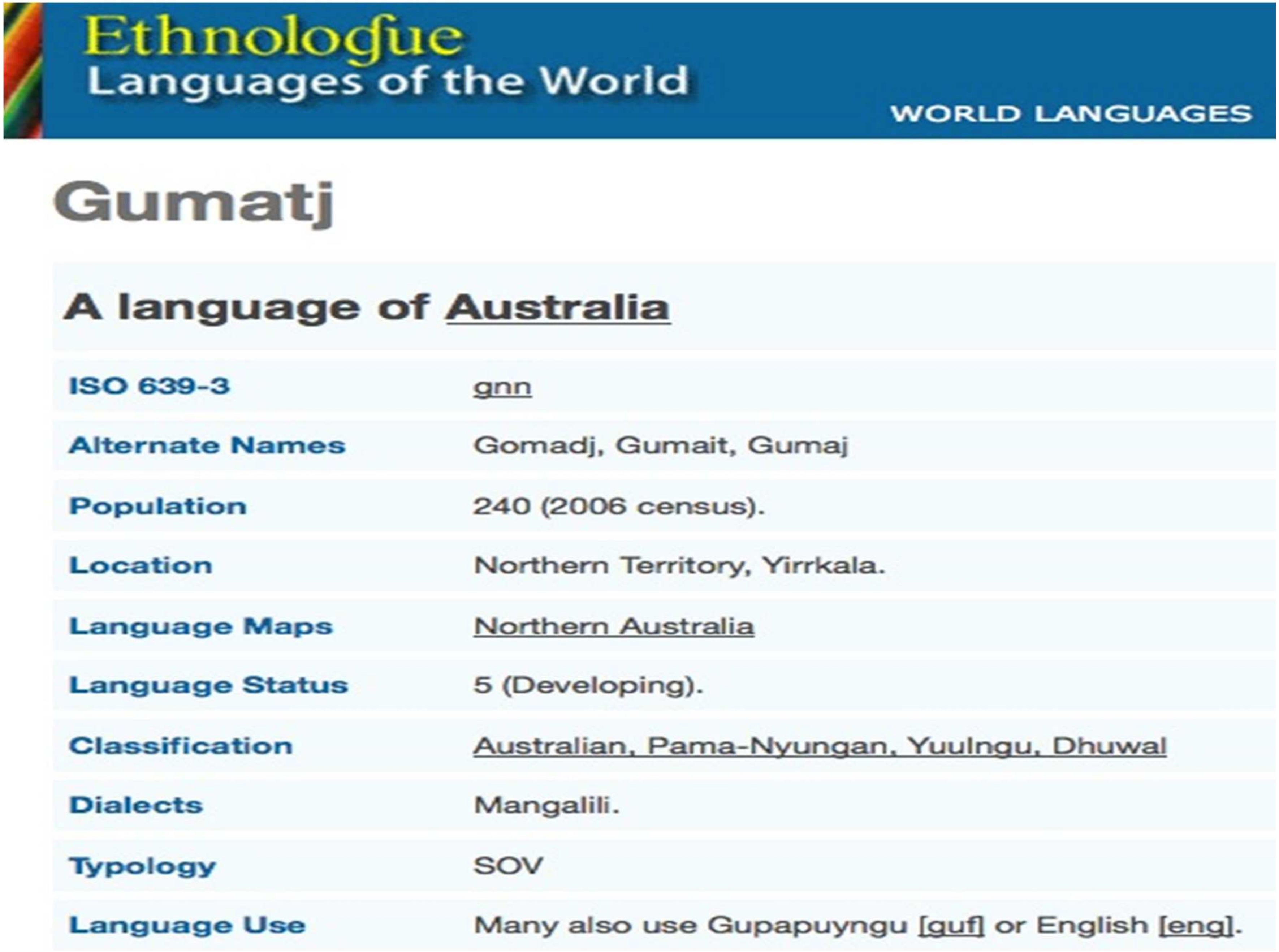


Djangu

A language of Australia

ISO 639-3	dhg
Alternate Names	Budalpudal, Burada, Buralbural, Buratha, Dangu, Dhangu, Dhangu'mi, Warameri, Waramiri, Warramiri, War-ramirri, Warumeri
Population	270 (2006 census). 58 Djangu, 170 Gaalpu, 44 Wangurri (2006 census).
Location	Northern Territory, Arnhem Land, Elcho island.
Language Maps	Northern Australia
Language Status	5 (Developing).
Classification	Australian, Pama-Nyungan, Yuulngu, Dhangu
Dialects	Dhangu-Djangu, Gaalpu (Kalbu), Golumala, Ngaymil, Rirratjingu, Wangurri.

http://www.ethnologue.com/language/gnn



Ethnologue
Languages of the World

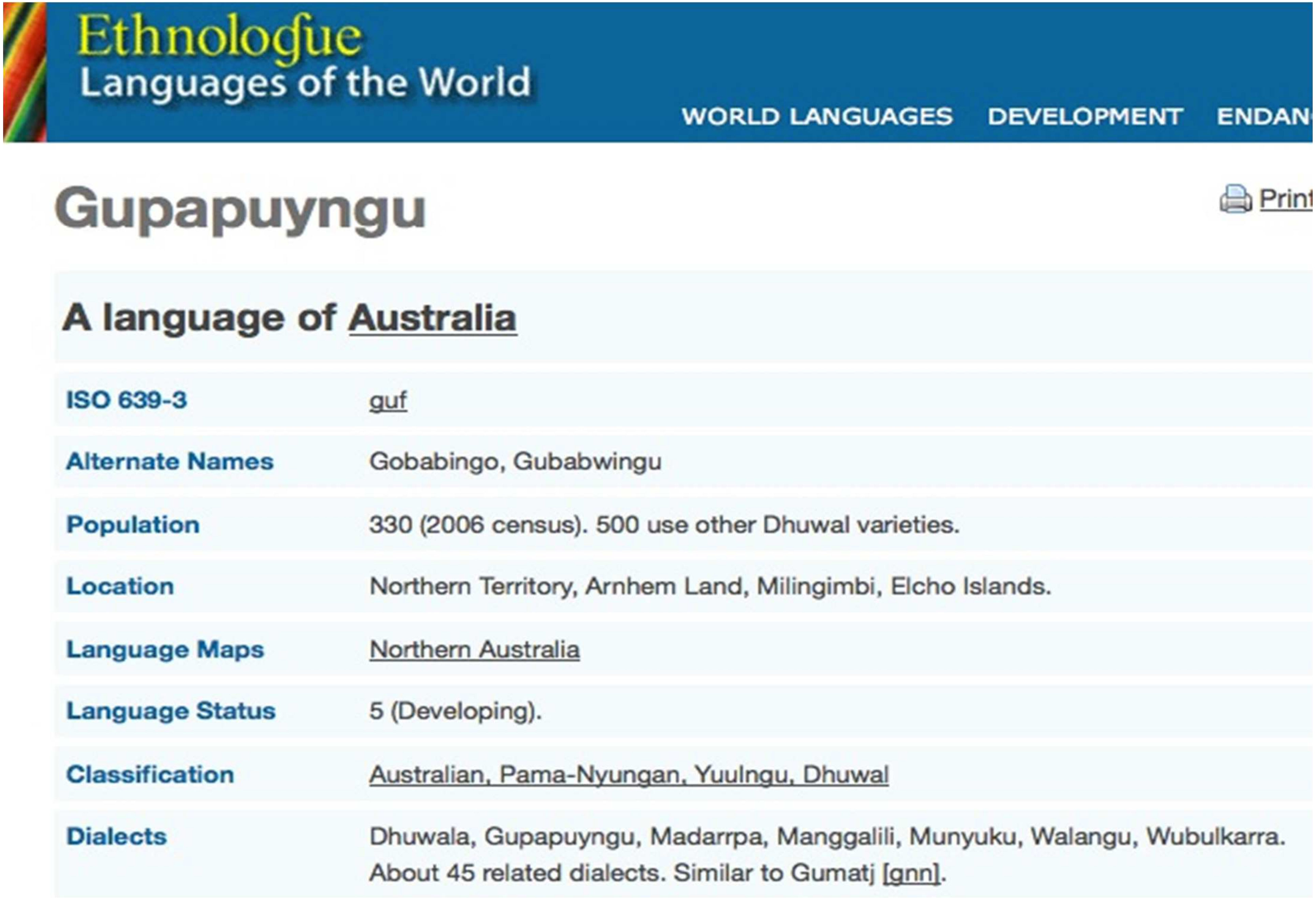
WORLD LANGUAGES

Gumatj

A language of Australia

ISO 639-3	gnn
Alternate Names	Gomadj, Gumait, Gumaj
Population	240 (2006 census).
Location	Northern Territory, Yirrkala.
Language Maps	Northern Australia
Language Status	5 (Developing).
Classification	Australian, Pama-Nyungan, Yuulngu, Dhuwal
Dialects	Mangalili.
Typology	SOV
Language Use	Many also use Gupapuyngu [guf] or English [eng].

http://www.ethnologue.com/language/guf

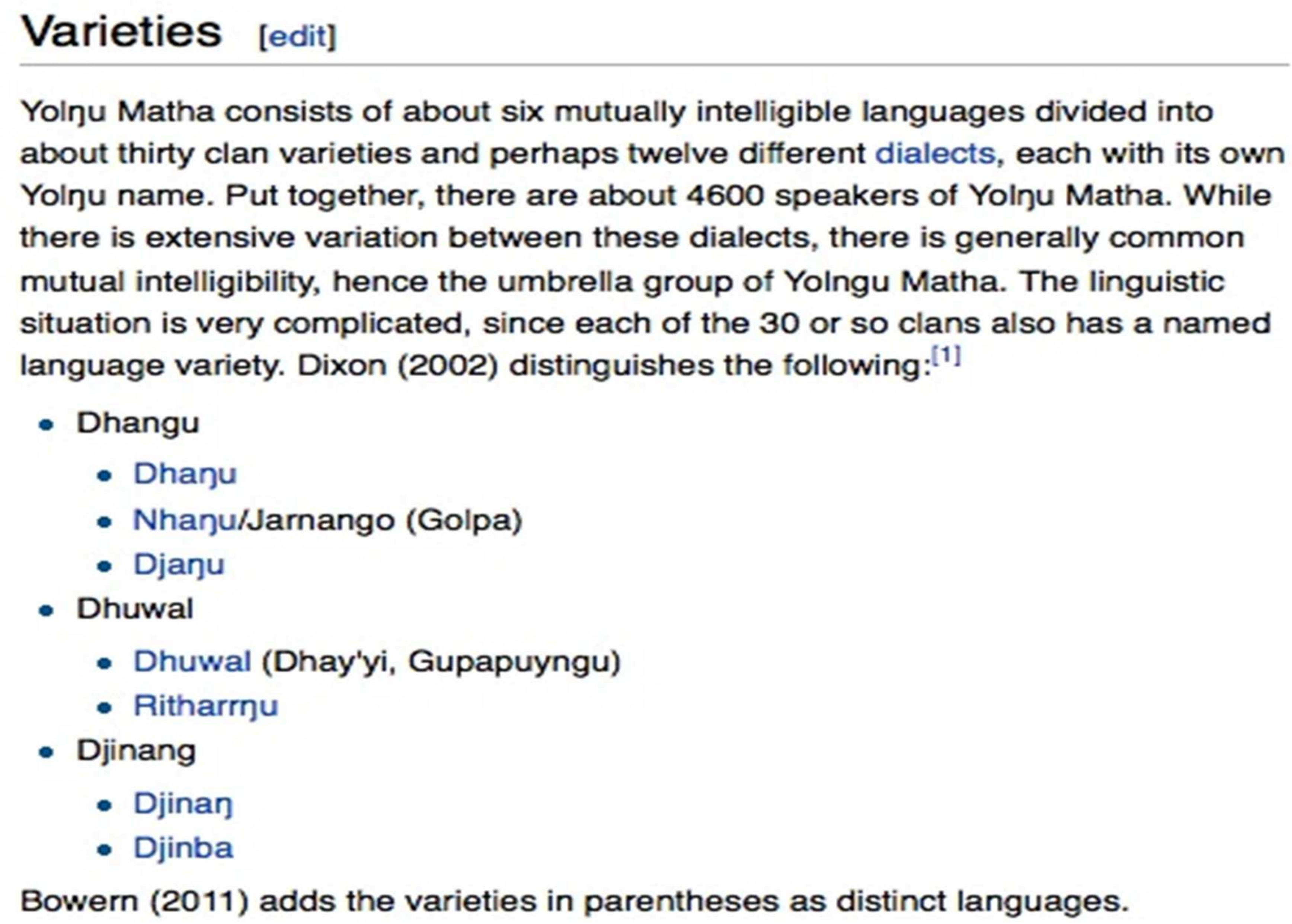


Gupapuyngu

A language of Australia

ISO 639-3	guf
Alternate Names	Gobabingo, Gubabwingu
Population	330 (2006 census). 500 use other Dhuwal varieties.
Location	Northern Territory, Arnhem Land, Milingimbi, Elcho Islands.
Language Maps	Northern Australia
Language Status	5 (Developing).
Classification	Australian, Pama-Nyungan, Yuulngu, Dhuwal
Dialects	Dhuwala, Gupapuyngu, Madarrpa, Manggalili, Munyuku, Walangu, Wubulkarra. About 45 related dialects. Similar to Gumatj [gnn].

Yolngu according to Wikipedia



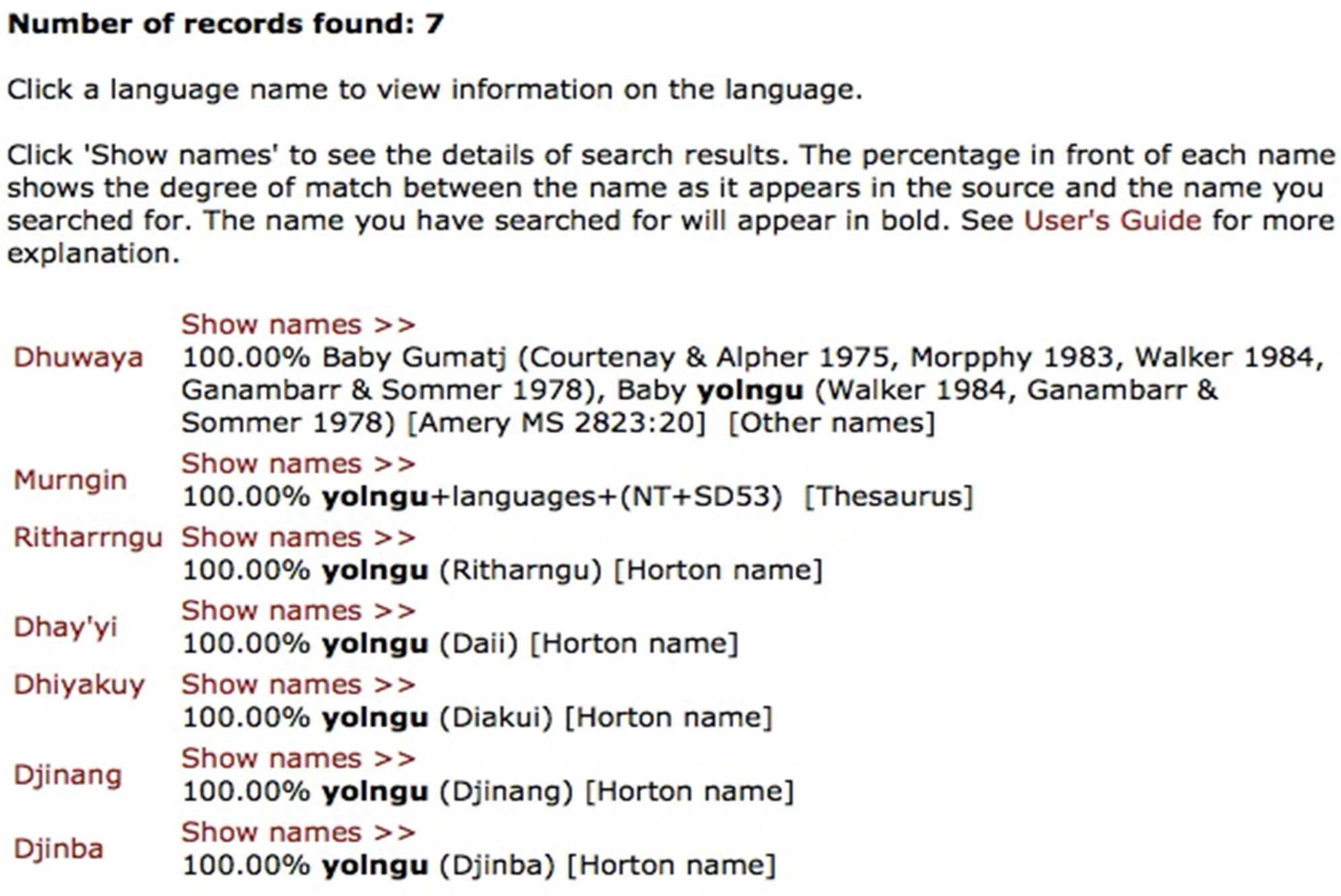
Varieties [edit]

Yolŋu Matha consists of about six mutually intelligible languages divided into about thirty clan varieties and perhaps twelve different dialects, each with its own Yolŋu name. Put together, there are about 4600 speakers of Yolŋu Matha. While there is extensive variation between these dialects, there is generally common mutual intelligibility, hence the umbrella group of Yolngu Matha. The linguistic situation is very complicated, since each of the 30 or so clans also has a named language variety. Dixon (2002) distinguishes the following:[1]

- Dhangu
  - Dhaŋu
  - Nhaŋu/Jarnango (Golpa)
  - Djaŋu
- Dhuwal
  - Dhuwal (Dhay'yi, Gupapuyngu)
  - Ritharrŋu
- Djinang
  - Djinaŋ
  - Djinba

Bowern (2011) adds the varieties in parentheses as distinct languages.

AustLang search - http://austlang.aiatsis.gov.au/main.php



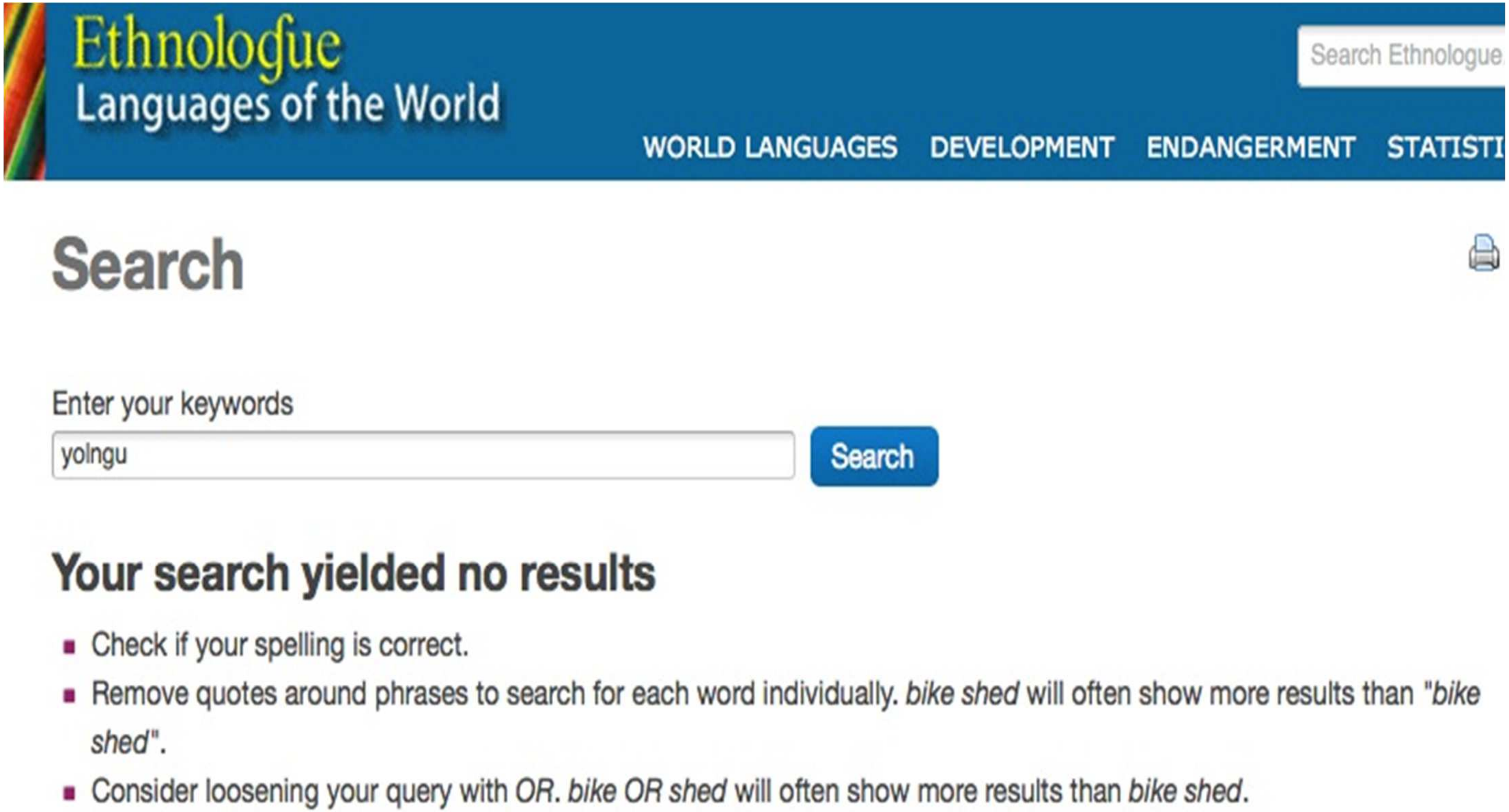
Number of records found: 7

Click a language name to view information on the language.

Click 'Show names' to see the details of search results. The percentage in front of each name shows the degree of match between the name as it appears in the source and the name you searched for. The name you have searched for will appear in bold. See User's Guide for more explanation.

Dhuwaya	Show names >> 100.00% Baby Gumatj (Courtenay & Alpher 1975, Morpphy 1983, Walker 1984, Ganambarr & Sommer 1978), Baby **yolngu** (Walker 1984, Ganambarr & Sommer 1978) [Amery MS 2823:20] [Other names]
Murngin	Show names >> 100.00% **yolngu**+languages+(NT+SD53) [Thesaurus]
Ritharrngu	Show names >> 100.00% **yolngu** (Ritharngu) [Horton name]
Dhay'yi	Show names >> 100.00% **yolngu** (Daii) [Horton name]
Dhiyakuy	Show names >> 100.00% **yolngu** (Diakui) [Horton name]
Djinang	Show names >> 100.00% **yolngu** (Djinang) [Horton name]
Djinba	Show names >> 100.00% **yolngu** (Djinba) [Horton name]

Yolngu flop!?!?



Search

Enter your keywords

yolngu

Your search yielded no results

- Check if your spelling is correct.
- Remove quotes around phrases to search for each word individually. *bike shed* will often show more results than "*bike shed*".
- Consider loosening your query with *OR*. *bike OR shed* will often show more results than *bike shed*.

Improving processes

- Different parts of ISO639 currently have different registration authorities
  - Part 1 – International Information Centre for Terminology
  - Part 2 – Library of Congress
  - Part 3 – SIL International
  - Part 5 – Library of Congress
  - Part 6 - Geolang
- Moving to a single registration authority would improve consistency of processes

Current developments

- ISO639-5 – language families and groups (2008), currently 114 codes
- ISO639-6 – comprehensive coverage of language variants (2009), four letter codes, number of codes assigned is not clear
- ISO639-4 - Implementation guidelines and general principles for language coding (most recent version 2010)

Current developments

- There are clear problems with parts 5 and 6
- Development of part 6 may have stalled
- Expert input is important for all parts
- But getting part 4 (general principles for language coding) as good as possible is very important

Influencing outcomes

- Representation on ISO Technical Committees is by national standards bodies
- Standards Australia has observer status only
- Discussion of issues happens at level of Working Groups
- Processes are opaque – difficult even to track membership of working groups

ARGILaRe

- Australian Reference Group for Interoperability of Language Resources
- Formed in February 2013
- Any interested people can join
- http://users.monash.edu.au/~smusgrav/ARGILaRe/

Research Data Alliance

- RDA:

  The purpose of the Research Data Alliance is to accelerate international data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonization, and discoverability.

Research Data Alliance

- RDA (rd-alliance.org):
  The purpose of the Research Data Alliance is to accelerate international data-driven innovation and discovery by facilitating research data sharing and exchange, use and re-use, standards harmonization, and discoverability.
- Improvements in identifying the language of resources is certainly relevant
- A Working Group within RDA is addressing problems for Standardisation of Categories and Codes

Research Data Alliance Overview

- Formed in 2012 by research funders from
  - EU (500M)
  - US (300M)
  - AU (23M)
- Members in every continent except Antarctica
  - But strongly biased towards US, Europe
- Plenary 2 took place in Washington DC in September 2013
  - https://www.rd-alliance.org/future-events
- Plenary 3 taking place in Dublin in March 2014
  - https://rd-alliance.org/rda-third-plenary-meeting.html

Working Groups and Interest Groups

- Interest Groups
  - people concerned with a particular class of problems (by discipline or by kind)
  - may have ongoing existence
  - will probably spin off series of working groups
- Working Groups
  - focussed on a particular problem
  - will run for 12-18 months
  - will produce a piece of infrastructure (broadly interpreted) for deployment
  - should lead to more data being exchanged

Mirror committee

- Standards Australia allows for participation: “in the work of international Technical Committees via a national mirror committee”
- The RDA Working Group is exploring the possibility of an Australian Mirror Committee for ISO TC 37
- A case for Net Benefit has to be made
  - This will require collaboration outside of research communities

Conclusion

- ISO639 has flaws but it is not going away
  - Would there ever be a right time to standardise?
- Efforts should be devoted to incremental improvement
  - Submit change requests
  - Join in efforts of groups like ARGILaRe and RDA
  - Find ways of collaborating with:
    - Various research communities
    - Interested parties outside academia