

# An Intelligent Search Infrastructure for Language Resources on the Web (SR0567353)

Chief Investigators: Timothy Baldwin, Steven Bird and Baden Hughes (University of Melbourne)



## Project Aims

**Build a language aware search engine**, using existing technologies and software components developed by the CI's, develop new infrastructure for language research.

## Context

Properties of the web in general:

- >10 billion web pages
- >7K languages

Specifically Australian interest in the multilingual web:

- 26% of Australians born overseas
- 38% of Australians speak language other than English as first language
- 62% of Australians speak a second language
- >500 Australian media organizations including the ABC and SBS deliver multilingual content in >130 languages

## Outcomes

**Language Crawler:** Selective web retrieval with emphasis on the languages of economic, scientific and cultural interest to Australia. Material stored in a centralized repository - indexed, annotated and preserved.

**Metadata Creation:** Each resource automatically classified, an Open Language Archives Community (OLAC) metadata record generated which identifies language and linguistic resource type. Texts annotated using robust language technologies to permit sophisticated indexing and retrieval.

**Language Archive:** High quality resources, including dictionaries and texts from endangered languages, archived to ensure ongoing accessibility. Archived content used as seed data for future crawls.

**Language-Aware Search Engine:** Permit users to enter language names, location names, linguistic constructions, or specify regions on a world map, returning URLs from crawled and archived content.

## Linkages

**National:** U. of Sydney, AIATSIS, State Library of Victoria, DSTC, CSIRO, National Library of Australia, PARADISEC

**International:** SIL International, U. Washington, Stanford U., U. Edinburgh, U. Arizona, California State U., Linguistic Data Consortium, U. of Pennsylvania, U. of Maryland, UMIACS, St Louis U.

## Collaboration Opportunities

We welcome engagement with:

- Creators, collectors and custodians of language data
- Representatives of language communities in Australia
- Researchers in web scale information retrieval

## Contact

E: [badenh@csse.unimelb.edu.au](mailto:badenh@csse.unimelb.edu.au)

