OLAC: Accessing the World's Language Resources

Gary Simons
SIL International and
Graduate Institute of Applied Linguistics

Steven Bird
University of Melbourne and
University of Pennsylvania

The Open Language Archives Community (OLAC) provides a standard for language resource description that permits users to simultaneously query the 40+ participating language archives in order to find resources among the 100,000+ items they hold. This poster demonstrates three mechanisms OLAC has developed for helping web users to discover language resources of potential interest within the collections of the participating archives.

Participating archives

The 43 participating institutions are listed on the left. Membership is open. If you know an institution whose materials should also be included, encourage them to join. For information on what this means, see the FAQs on the OLAC home page (www.language-archives.org).

Aboriginal Studies Electronic Data Archive (ASEDA)
Australian Institute of Aboriginal and Torres Strait Islander Studies, Canberra, Australia

Academia Sinica Collections
Institute of Linguistics, Academia Sinica, Taipei, Taiwan

African Language Materials Archive
The West African Research Association, the Council of American Overseas Research Centers, Columbia University Libraries--African Studies, and the Information Society Division of the United Nations Educational, Scientific and Cultural Organization, n/a

Alaska Native Language Archive
University of Alaska Fairbanks, Fairbanks, Alaska

Archive of the Indigenous Languages of Latin America
University of Texas at Austin, Austin, USA

ATILF Resources
ATILF, Nancy, France

Audio Archive of Linguistic Fieldwork
University of California, Berkeley Language Center, Berkeley, CA, USA

Boiste
OLAC, Philippe KREBS, Paris, FRANCE

C'ek'aedi Hwnax Ahtna Regional Linguistic and Ethnographic Archive
Ahtna Heritage Foundation, Copper Center, Alaska, United States of America

Central Institute of Indian Languages: Publications
Central Institute of Indian Languages, Mysore, India

Centre de Ressources pour la Description de l'Oral (CRDO)
CNRS / Centre de Ressources pour la Description de l'Oral (CRDO), Paris, France

CHILDES Data repository
Carnegie Mellon University, Pittsburgh, USA

Comparative Corpus of Spoken Portuguese
IEL Unicamp, Campinas - SP - Brasil

Cornell Language Acquisition Laboratory, CLAL
Cornell Language Acquisition Lab (CLAL), Ithaca, New York

CRDO archive
Resource Center for the Description of Oral - Centro de Recursos para la Descripción del Oral - Centre de ressources pour la description de l'oral (CRDO), France

A Digital Archive of Research Papers in Computational Linguistics
Philadelphia, USA

ELRA Catalogue of Language Resources
ELRA (European Language Resources Association), Paris, France

Ethnologue: Languages of the World
SIL International, Dallas, USA

Graduate Institute of Applied Linguistics Library
Graduate Institute of Applied Linguistics (GIAL), Dallas, TX

IMDI to OAI bridge
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

Kaipuleohone
University of Hawaii at Manoa, Honolulu, USA

Language Commons Language Corpora
Language Commons, San Francisco, USA

The LDC Corpus Catalog
LDC, Philadelphia, USA

The LINGUIST List Language Resources
The LINGUIST List, Ypsilanti, MI

Magoria Books' Carib and Romani Archive
Magoria Books, Online

MMM - Multi-Modal Media File Server
Idiap Research Institute, Martigny, Switzerland

Multimodal Learning and teaching Corpora Exchange
Universite de Franche-Comte, Besancon, France

The Natural Language Software Registry
DFKI, Saarbrucken, Germany

ODIN - The Online Database of Interlinear Text
California State University, Fresno, Fresno, CA, USA

Oxford Text Archive
Oxford Text Archive, Oxford, UK

Pacific And Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)
A consortium made up of the University of Melbourne, University of Sydney, and the Australian National University, Melbourne, Sydney, Canberra, Australia

Perseus Digital Library
Tufts University, Somerville, USA

The Rosetta Project: A Long Now Foundation Library of Human Language
Long Now Foundation, San Francisco, CA

SIL Language and Culture Archives
SIL International, Dallas, USA

Surrey Morphology Group Databases
Surrey Morphology Group, University of Surrey, Guildford, UK

Survey of California and Other Indian Languages
University of California, Berkeley, Department of Linguistics, Berkeley, CA, USA

TALKBANK Data repository
Carnegie Mellon University, Pittsburgh, USA

Tibetan and Himalayan Digital Library
University of Virginia, Charlottesville, USA

The Typological Database Project
The Typological Database Project, Utrecht, Netherlands

U Bielefeld Language Archive
University of Bielefeld, Bielefeld, Germany

Virtual Kayardild Archive
University of Melbourne, Melbourne, Australia

WALS Online
Max Planck Digital Library, Munich, Germany

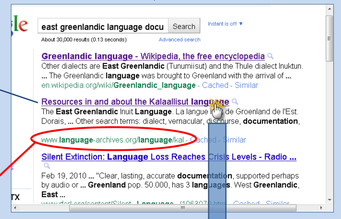
WALS Online RefDB
Max Planck Digital Library, Munich, Germany

Three ways to find resources

1. Google search

Type a language or dialect name and resource type (e.g. dictionary, grammar) into a Google search box.

For this query, OLAC's index page for this language is the second hit. For uncommon languages, the OLAC index page is typically high.



The ISO 639-3 code is kal, so the page URL is:
http://www.language-archives.org/language/kal

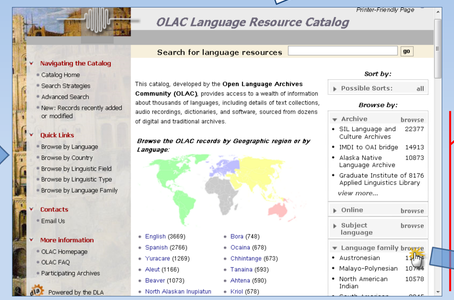
Substitute any valid ISO 639-3 code at the end.

The page lists all known alternate names and dialect names; therefore searches can use any of these names, not just ISO's official reference name.

2. Browse by country and language from www.language-archives.org

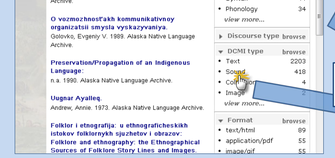
Via the Greenland country page

3. Explore via faceted search at search.language-archives.org

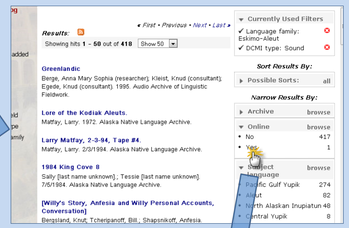


Click to list all language families (from ISO 639-5).

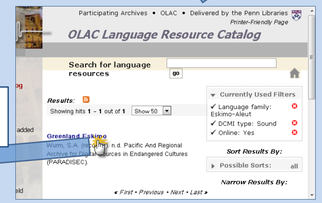
Click to list the 3,077 items for the Eskimo-Aleut family.



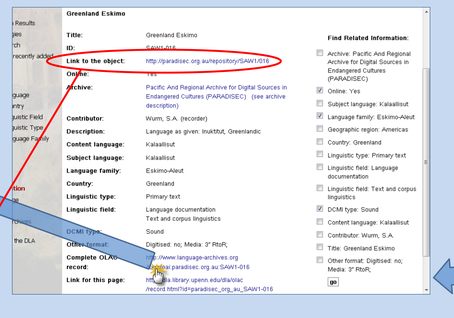
Click to list the 418 items that are Sound.



Click to list the single item that is online.



Click title for full details.

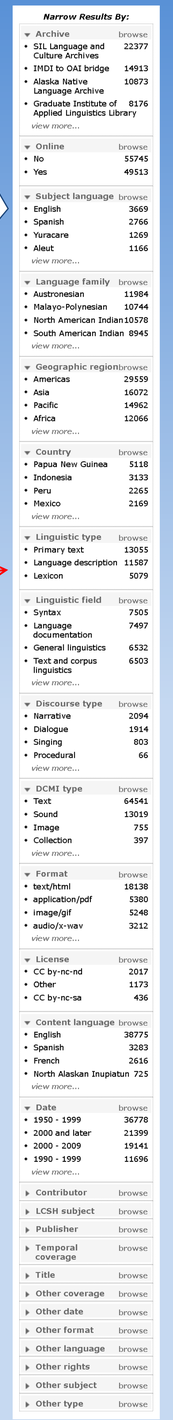


Access this online resource directly

Flexible search using facets

The complete list of 26 facets is listed on the right. The first 14 facets are opened to show the four most frequent values. They come from controlled vocabularies that OLAC uses to permit consistent search across all the participating archives. The final 12 facets have freeform content.

The search interface always shows the current facet list in the right pane. Following each facet value is the number of items in current selection which have that value. At the outset, all 100,000+ items are selected. Clicking on a facet value narrows the selection to just the items with that value and causes the list of values with counts to be immediately recalculated to reflect the items in the new, more narrow selection.



Acknowledgments: The work reported in this poster has been supported by a collaborative grant from the National Science Foundation awarded to the University of Pennsylvania (BCS- 0723357) and to the Graduate Institute of Applied Linguistics (BCS-0723864). We are deeply indebted to Haejoong Lee and Delphine Khanna who were primary implementers for the services presented here.