National Science Foundation
Workshop for
Documenting Endangered Languages

October 14 - 17, 2007
The New England Center,
University of New Hampshire,
Durham, NH

All materials contained herein are for viewing only. Reproduction or dissemination without permission is strictly prohibited. For permission or more information, write to DEL@NSF.GOV
An Infrastructure for Indexing Endangered Language Documentation

Gary F. Simons  
*SIL International and Graduate Institute of Applied Linguistics*

DEL Workshop, Durham, NH, 15-16 Oct 2007
What is OLAC?

www.language-archives.org

► OLAC is an international partnership of institutions and individuals who are creating a world-wide virtual library of language resources by:
  ▪ Developing consensus on best current practice for the digital archiving of language resources
  ▪ Developing a network of interoperating repositories and services for housing and accessing such resources

► Founded in December 2000
  ▪ Now has 34 participating archives
Who’s involved?

- Aboriginal Studies Electronic Data Archive
- Academia Sinica
- Alaska Native Language Center
- Archive of Indigenous Languages of Latin America
- ATILF Resources
- Berkeley Language Center
- Centre de Ressources pour la Description de l'Oral
- CHILDES Data Repository
- Comparative Corpus of Spoken Portuguese
- Cornell Language Acquisition Laboratory
- Dictionnaire Universel Boiste 1812
- DOBES catalogue (MPI, Nijmegen)
- Ethnologue: Languages of the World
- European Language Resources Association
- Laboratoire Parole et Langage
- Linguistic Data Consortium Corpus Catalog
- LINGUIST List Language Resources
- Natural Language Software Registry
- Online Database of Interlinear Text (ODIN)
- Oxford Text Archive
- PARADISEC
- Perseus Digital Library
- Research Papers in Computational Linguistics
- Rosetta Project 1000 Language Archive
- SIL Language and Culture Archives
- Surrey Morphology Group Databases
- Survey for California and Other Indian Languages
- TalkBank
- Tibetan and Himalayan Digital Library
- TRACTOR
- Typological Database Project
- University of Bielefeld Language Archive
- University of Queensland Flint Archive
- Virtual Kayardild Archive (Melbourne)
How does it work?

► The 34 archives submit catalogs in a standard form …

► which supplies information to search services.

► to the OLAC aggregator …

Linguist List, EMU

LDC, U Penn

Gary Simons
SIL

NSF Documenting Endangered Languages Workshop,
Durham, New Hampshire, October 2007
Results from "ethnologue.com"

1. ***** oai:ethnologue.com:POT Similar records by: score date
title: POTAWATOMI: a language of USA
description: A page from the Web edition of Ethnologue: Languages of the World (14th edition) giving basic facts about the language and where it is spoken.

Results from "linguistlist.org"

1. ***** oai:linguistlist.org:lang POT Similar records by: score date
title: LINGUIST List Resources for Potawatomi
description: A page listing all resources ...

Results from "sil.org" List all results from this archive (2 matches)

1. ** oai:sil.org:11119 Similar records by: score date subject
title: Patterns of person-number reference in Potawatomi
description: http://www.ethnologue.com/show_work.asp?id=11119
subject: Reference

Results from "perseus.tufts.edu" List all results from this archive (5 matches)

description: Descriptions of the Potawatomi, Miami, Sauk, Menomone [Menominee], Winnebago, and Dacota [Sioux] provide insights about the observers as well as the peoples observed.
title: Narrative of an expedition to the source of St. Peter's River, Lake Winnipeek, Lake of the Woods, &c. &c. performed in the year 1823, by order of the Hon. J.C. Calhoun, Secretary of war, under the command of Stephen H.
### What is the current coverage?

<table>
<thead>
<tr>
<th></th>
<th>All archives</th>
<th>Excluding \textit{Ethnologue}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items in catalog</td>
<td>30,591</td>
<td>23,292</td>
</tr>
<tr>
<td>ISO 639-3 languages included</td>
<td>7,299</td>
<td>3,134</td>
</tr>
<tr>
<td>Items that are online</td>
<td>16,018 (52%)</td>
<td>8,719 (37%)</td>
</tr>
</tbody>
</table>
## OLAC coverage in relation to language size

<table>
<thead>
<tr>
<th>Population range</th>
<th>Languages</th>
<th>In OLAC</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000,000 or more</td>
<td>83</td>
<td>81</td>
<td>98%</td>
</tr>
<tr>
<td>1,000,000 to 9,999,999</td>
<td>264</td>
<td>209</td>
<td>79%</td>
</tr>
<tr>
<td>100,000 to 999,999</td>
<td>892</td>
<td>507</td>
<td>57%</td>
</tr>
<tr>
<td>1,000 to 99,999</td>
<td>3,746</td>
<td>1,593</td>
<td>43%</td>
</tr>
<tr>
<td>100 to 999</td>
<td>1,071</td>
<td>356</td>
<td>33%</td>
</tr>
<tr>
<td>1 to 99</td>
<td>548</td>
<td>237</td>
<td>43%</td>
</tr>
<tr>
<td>Unknown</td>
<td>308</td>
<td>56</td>
<td>18%</td>
</tr>
<tr>
<td><strong>All living languages</strong></td>
<td>6,912</td>
<td>3,039</td>
<td>44%</td>
</tr>
<tr>
<td><strong>Extinct languages</strong></td>
<td>95</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Online resources in relation to language size

<table>
<thead>
<tr>
<th>Population range</th>
<th>Languages</th>
<th>In OLAC</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>10,000,000 or more</td>
<td>83</td>
<td>75</td>
<td>410</td>
</tr>
<tr>
<td>1,000,000 to 9,999,999</td>
<td>264</td>
<td>170</td>
<td>297</td>
</tr>
<tr>
<td>100,000 to 999,999</td>
<td>892</td>
<td>326</td>
<td>467</td>
</tr>
<tr>
<td>1,000 to 99,999</td>
<td>3,746</td>
<td>859</td>
<td>1,054</td>
</tr>
<tr>
<td>100 to 999</td>
<td>1,071</td>
<td>130</td>
<td>179</td>
</tr>
<tr>
<td>1 to 99</td>
<td>548</td>
<td>95</td>
<td>146</td>
</tr>
<tr>
<td>Unknown</td>
<td>308</td>
<td>22</td>
<td>28</td>
</tr>
</tbody>
</table>

### All living languages

<table>
<thead>
<tr>
<th>Languages</th>
<th>In OLAC</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,912</td>
<td>1,677</td>
<td>2,581</td>
</tr>
</tbody>
</table>

### Extinct languages

<table>
<thead>
<tr>
<th>Languages</th>
<th>In OLAC</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td></td>
<td>42</td>
</tr>
</tbody>
</table>
Conclusion

► OLAC has a functioning infrastructure that allows our community to index and discover endangered language documentation.

► We need to encourage every institution with endangered language resources to participate so that the catalog can be complete.

► We could enhance standards for resource description to support metrics for summarizing the extent of documentation for a language.