National Science Foundation
Workshop for
Documenting Endangered Languages

October 14 - 17, 2007

The New England Center,

University of New Hampshire,
Durham, NH

OLAC: Open Language Archives Community

OPEN LANGUAGE ARCHIVES

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
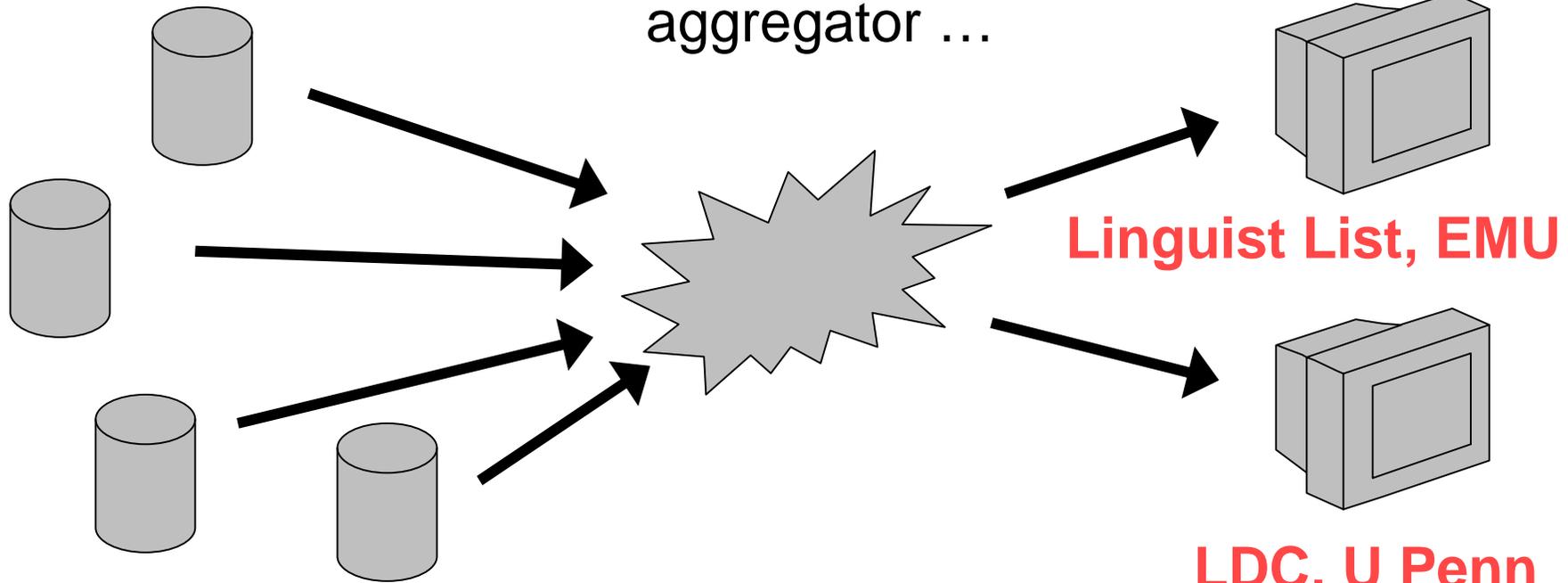- ► Virtual Kayardild Archive (Melbourne)

**Gary Simons
SIL**

NSF Documenting Endangered Languages Workshop, Durham, New Hampshire, October 2007

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**

What is the current coverage?

	All archives	Excluding *Ethnologue*
Items in catalog	30,591	23,292
ISO 639-3 languages included	7,299	3,134
Items that are online	16,018 (52%)	8,719 (37%)

Gary Simons
SIL

NSF Documenting Endangered Languages Workshop, Durham, New Hampshire, October 2007

OLAC coverage in relation to language size

Population range	Languages	In OLAC		Items
10,000,000 or more	83	81	98%	2,533
1,000,000 to 9,999,999	264	209	79%	1,281
100,000 to 999,999	892	507	57%	2,097
1,000 to 99,999	3,746	1,593	43%	8,131
100 to 999	1,071	356	33%	1,866
1 to 99	548	237	43%	704
Unknown	308	56	18%	167
All living languages	6,912	3,039	44%	16,779
Extinct languages		95		231

Online resources in relation to language size

Population range	Languages	In OLAC		Items
10,000,000 or more	83	75	90%	410
1,000,000 to 9,999,999	264	170	64%	297
100,000 to 999,999	892	326	37%	467
1,000 to 99,999	3,746	859	23%	1,054
100 to 999	1,071	130	12%	179
1 to 99	548	95	17%	146
Unknown	308	22	7%	28
All living languages	6,912	1,677	24%	2,581
Extinct languages		33		42

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.

Gary Simons
SIL

NSF Documenting Endangered Languages Workshop, Durham, New Hampshire, October 2007