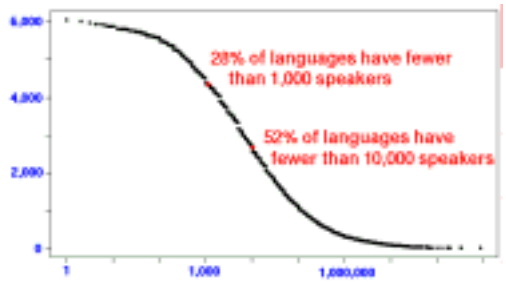
Open Language Archives

Steven Bird, University of Pennsylvania
Gary Simons, SIL International

1

The World's Languages



2

Countries with >150 languages

New Guinea:	**823**	Australia:	**235**
Indonesia:	**726**	Congo (DRC):	**218**
Nigeria:	**505**	China (PRC):	**201**
India:	**387**	Brazil:	**192**
Mexico:	**288**	USA:	**176**
Cameroon:	**279**	Philippines:	**169**

3

Major Language Archives

- **American Philosophical Society**
  - Wordlists, texts, manuscripts, audio; 200 languages
- **National Anthropological Archives**
  - manuscripts, field-notes, photographs, maps, video
  - 1,300 recordings of myths, legends, stories, songs
- **Perseus Project**
  - >70 million words of Greek, Latin, English, Italian, German
- **Aboriginal Studies Electronic Data Archive**
  - texts, dictionaries, grammars and teaching materials
  - 300 Australian languages

4

Major European Archives

- **Germany**
  - **IDS**: Institüt für Deutsche Sprache (Mannheim)
  - **BAS**: Bavarian Archive of Speech (Munich)
- **France**
  - **INALF**: Institute National à Langue Français (Paris)
  - **LACITO**: Langues et Cultures à Tradition Orale (Paris)
- **United Kingdom**
  - **OTA**: Oxford Text Archive (Oxford)
- **Many others …**

5

Alaska Native Language Center

- Founded in 1972
- 20 native languages
- 10,000 documents
  - Texts
  - Ethnographies
  - Place names
  - Lexicons
- 3,000 recordings

6

An ANLC Record

Title:	Gwich'in Wordlist ← RESOURCE TYPE?
Author:	Zimmerman, Herbert
Date:	1959
Language:	Gwich'in ← LANGUAGE NAME?
Format:	Non-digital ← AVAILABILITY?
Description:	MS, 75pp
Description:	1400 items based on SIL schedule

7

American Indian Studies Research Institute, Indiana

- Interactive language lessons for American Indian languages
- Multimedia dictionaries
  - audio
  - photographic images

8

UC Berkeley Survey of Californian Languages

- 90 languages
- Field notes
- 750 cassettes
- Catalog is an HTML document
- Typical…

9

Linguistic Data Consortium

- **Data for new language technologies:**
  - ASR, NLP, MT, IR, TREC, MUC, TDT, …
  - ~200 CD-ROM publications (largest 82 CDs)
  - >1 terabyte of audio data
- **E.g. SWITCHBOARD Corpus**
  - 2400 transcribed telephone calls
  - Distributed on 26 CDs (web is inappropriate)
  - Published, ISBN, distribution mechanism

10

ACL Natural Language Software Repository

- Hosted by the German Foundation for AI (DFKI)
- Software metadata:
  - Authors
  - Functionality
  - Linguistic datatype (e.g. lexicon)
  - File format
  - Operating system
  - availability
  - URL

11

Taking Stock: Resource Types

- **DATA**
  - Sound recording
  - Shoebox of hand-written index cards
  - Descriptive grammar
- **TOOLS**
  - Software for creating, storing, querying and viewing language data
  - Formats for storage and interchange (e.g. TEI)
- **ADVICE**
  - Mailing list archives, FAQs

12

Taking Stock: The Community

- Linguists
  - >13,000 members of LINGUIST
  - Ethnologue >500,000 page hits / month
- Engineers
  - ~1,000 organizations which buy LDC resources
- Language teachers
- Archivists
- Software developers

13

Challenges

- Endangered languages
  - Preserving languages before they die
- Endangered data
  - Saving old recordings before they disintegrate
- Best practices
  - Creating new data using XML and Unicode
- Finding aids
  - Locating resources (mailing lists)

14

Finding Aids

- Goal: “bringing like things together and differentiating among them” (Svenonius)
- Traditional databases versus the web
  - Metadata is coherent, but highly distributed
- We need a middle ground:
  - Bottom-up, distributed initiatives
  - Consistent, centralized finding aids

15

Language Archives within the OAI

- Specialist communities can define their own metadata format
- Service providers can exploit the metadata
- Philadelphia Workshop (December 2000)
  - linguists, anthropologists, archivists, engineers, funding agencies, publishers
  - North America, South America, Europe, Middle-East, Africa, Asia, Australia
  - Commitment to implement OAI

16

Structure of OLAC

Three groups:
- Advisory board
- Member archives
- Participating data providers

Three phases:
- Alpha test [Dec 2000]
- Pilot [Fall 2001]
- Operational [Fall 2002]

17

Primary Service Provider

- Eastern Michigan Univ & Wayne State Univ
- Funded by NSF
- >13,000 members
- Complete union catalog

18

A Community defined by its metadata

OPEN
- Rights.openness
- Format.openness

LANGUAGE
- Encoding scheme: RFC 1766
- Subject.language

ARCHIVES
- Type.data
- Type.functionality

19

Language Identification

- Existing standards (ISO 639, RFC 1766)
  - incomplete: 7% coverage
  - inconsistent: e.g. Quechua, Bantu (other)
  - Undocumented: only gives a name
- Issues to be addressed:
  - Impossible to create a static inventory
  - Multiple names for a language
    - E.g. ANLC: Gwich'in versus Kutchin

20

SIL Ethnologue

- The only complete language identification scheme openly available on the web
- For each of 6,800 languages:
  - Language name and variants, 3-letter code
  - Population, location
  - Linguistic classification
  - Dialects, alternative names for dialects
  - Notes on language use and available literature

21

Progress on Data Providers

- Linguistic Data Consortium
- European Language Resources Assocation
- German Foundation for AI (DFKI)
- SIL International
- Perseus Project
- Alaska Native Language Center
- LACITO
- CBOLD: Comparative Bantu Online Lexical Database

22

LDC Prototype Service Provider

Harvests data from LDC, ELRA, DFKI

Query for "language=Bulgarian":

oai:ldc:LDC95T5	ECI Multilingual Text Lang: Albanian, **Bulgarian**, Chinese, Czech, … Applications: IR, MT, LM
oai:elra:L0030	Bulgarian Morphological Dictionary Lang: **Bulgarian** 67,500 entries, 242 inflectional types, …
oai:dfki:KPML	Grammar development workbench Lang: Spanish, Russian, Japanese, **Bulgarian**, …

23

Our Experience with the OAI

- Experience of OLAC alpha testers
  - Harvesting protocol
  - Dublin Core
- OAI support
  - Specialized metadata
  - OAI representative at our meeting (Michael Nelson)
  - Solves our problem with cataloging distributed, dynamic resources

24

Challenges ahead…

- Large legacy catalogs
  - cleansing and exporting
  - hierarchical collections
- Overlap with other OAI groups
  - e-prints & digital museums
- OAI as a springboard
  - digitization of legacy data
  - formats for access in perpetuity

25