The Open Language Archives Community:

Building a worldwide library of digital language resources

Gary Simons, SIL International

LSA Tutorial on *Archiving and Linguistic Resources*

6 Jan 2005, Oakland, CA

Unprecedented opportunity

- Digital archiving of language documentation and description on the World-Wide Web offers:
  - Minimal cost multimedia publishing
  - Maximal access by the citizens of the world
- This holds the promise of unparalleled access to information.

Or, Unprecedented chaos?

- Pursuing digital archiving of language documentation in isolation will result in:
  - Resources that are as good as lost since others won't be able to find them.
  - Resources that are not usable by others due to the proliferation of idiosyncratic formats and practices.
- This holds out the specter of unparalleled frustration and confusion.

The vision

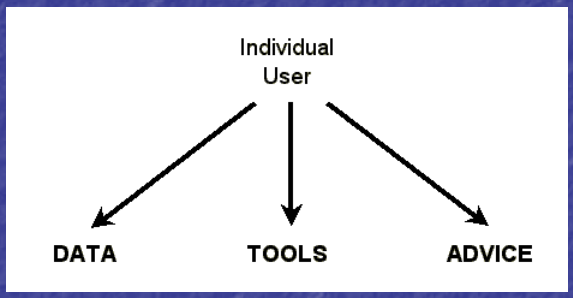
- Fulfill the promise (and avoid the specter) by acting in community to define and follow best common practice
- A gap analysis:
  - What users want—the ideal
  - What users actually get —the gap
  - What it would take to bridge the gap—a community infrastructure

What users want

The individuals who use and create language documentation and description are looking for three things:

- Primary and secondary data about languages
- Computational tools to create, view, query, or otherwise use language data
- Advice on how best to do the above

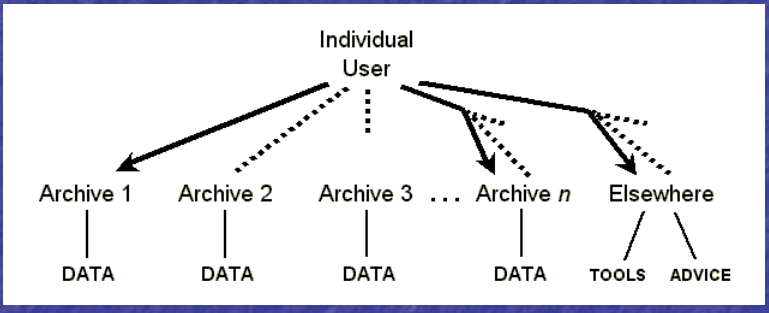
The ideal situation



What users actually get

- The data are archived at hundreds of sites
  - Some are on Web and user finds them
  - Some are on Web but user can't find them
  - Some are not even on Web
- The tools and advice are at different sites than the data

The gap



It's even worse

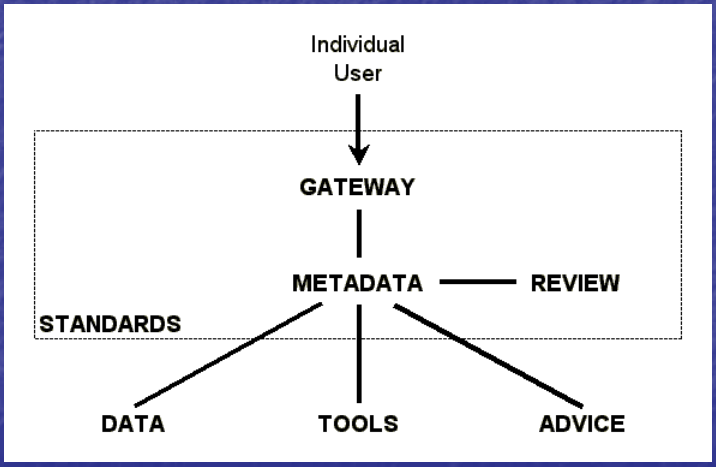
- The user may not find all existing data about the language of interest because different sites have called it by different names.
- The user may not be able to use an accessible data file for lack of being able to match it with the right tools.
- The user may locate advice that seems relevant but then has no way to judge how good it is.

What a community could provide

In order to bridge the gap, the individuals who use and create language documentation and description need a community with standards that define:

- Uniform metadata for describing resources
- A single gateway for finding resources
- A process to review practices and standards

A community infrastructure



Open Language Archives Community

OLAC is an international partnership of institutions and individuals who are creating a worldwide 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

Participating Archives

- Aboriginal Studies Electronic Data Archive (ASEDA)
- Academia Sinica
- Alaska Native Language Center
- Archive of Indigenous Languages of Latin America (AILLA)
- ATILF Resources
- CHILDES Data Repository
- Cornell Language Acquisition Laboratory (CLAL)
- Dictionnaire Universel Boiste 1812
- Digital Archive of Research Papers in Computational Linguistics
- Ethnologue: Languages of the World
- European Language Resources Association (ELRA)
- LACITO Archive
- LDC Corpus Catalog
- LINGUIST List Language Resources
- Natural Language Software Registry
- Oxford Text Archive
- PARADISEC
- Perseus Digital Library
- Rosetta Project 1000 Languages
- SIL Language & Culture Archives
- Surrey Morphology Group Databases
- Survey for California and Other Indian Languages
- TalkBank
- Tibetan and Himalayan Digital Library
- TRACTOR
- Typological Database Project
- Univ. of Bielefeld Language Archive
- Univ. of Queensland Flint Archive

Metadata standard

- Based on Dublin Core metadata standard:
  - Contributor, Coverage, Creator, Date, Description, Format, Identifier, Language, Publisher, Relation, Rights, Source, Subject, Title, Type
- OLAC adds extensions (with controlled vocabularies) specific to our community:
  - Language Identification, Linguistic Data Type, Linguistic Field, Participant Role, Discourse Type

Gateway standard

- Based on a Digital Library Federation standard
  - Open Archives Initiative Protocol for Metadata Harvesting
  - *Service providers* use the protocol to harvest metadata from *data providers*
- OLAC has four ways to become a data provider
  - Implement a dynamic interface to existing database
  - Map existing database to a static XML document
  - Use web forms of OLAC Repository Editor service
  - *Under development:* Install an E-prints server

Process standard

- Defines how OLAC is organized:
  - Coordinators, Advisory Board, Council, Archives, Services, Working Groups, Participating individuals
- Defines three types of documents:
  - Standards, Recommendations, Notes
- Defines how a document moves from one life-cycle status to another.
  - Draft, Proposed, Candidate, Adopted, Retired

OPEN LANGUAGE ARCHIVES

Call for participation

- All institutions and individuals with language resources to share are enthusiastically invited to participate.
- Visit **www.language-archives.org** to:
  - Try our two search services
  - Read workpapers and published articles
  - Subscribe to the OLAC-General mailing list
  - Learn how to become a data provider