Building cyberinfrastructure for the language acquisition community

The role of the Open Language Archives Community (OLAC)

Gary F. Simons
SIL International

14 July 2007

Applying Cyberinfrastructure to the Language Sciences: A Case Study for Language Acquisition
Stanford University, Palo Alto, CA
Two possible futures

Digital archiving could hold:

► the promise of unparalleled access to information,
  or,
► the specter of unparalleled frustration and confusion

The outcome will depend on whether we:

► act in community to define and follow best practice,
  or,
► act in isolation to proliferate idiosyncratic practices
Open Language Archives Community

www.language-archives.org

- 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

- Founded in December 2000
  - Now has 34 participating archives
Participating Archives

- 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'Oural
- 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)
Interoperation

■ Definition

▲ Interoperability is the ability for two or more systems to exchange information or services and to make satisfactory use of what is exchanged.

■ Two levels of interoperation

▲ Macrointeroperation: interoperating at the level of resource discovery

▲ Microinteroperation: interoperating within the content of relevant resources
What linguists want

- Estimated size of today’s public Internet:
  - 100 million distinct web sites
  - 30 billion unique web pages

- “The language resource I’m looking for is probably out there somewhere, but …”
  - Which site should I look on?
  - How do I find the exact resource once I arrive?
  - It’s like searching for a needle in a haystack!

- The general strategy
  - Aggregate and Filter
Aggregate and Filter

The value chain of Internet publishing:

- Aggregators solve the “Which site?” problem by creating a single place to look.
- Filters address the “How do I find it?” problem by showing only the resources that match your criteria.
Two kinds of interoperation

- **Shallow interoperation**
  - Generic to all problem domains
  - Aggregates everything reachable via the ubiquitous HTTP infrastructure
  - Filters on the surface content of plain text

- **Deep interoperation**
  - Built for a specific problem domain
  - Uses domain-specific protocol to aggregate only what is relevant to the community
  - Uses domain-specific markup and vocabularies to filter on underlying concepts and structures
Good news and bad news

- Regarding shallow interoperation
  - It exists on a global scale (e.g. Google, Yahoo!) and is easy to support and use
  - But it gives poor results for language resources
    - Lots of noise (= low precision): words used in queries have many irrelevant senses
    - Lots of drop out (= low recall): words used in queries have synonyms and translations

- Regarding deep interoperation
  - It gives both high recall and high precision
  - But it takes more work to follow standards
Anchored by an aggregator

1. The community needs an aggregator for language resources to anchor its cyberinfrastructure.

- Provides a single authoritative inventory of every resource in the treasury of linguistic knowledge
- This becomes the basis for resource discovery within the community
- An open web services API allows members of the community to build services that add value for the community
2. The community needs a metadata standard that describes the aggregated resources in a way that will support its filtering needs.

- The digital library community has developed a generic descriptive metadata standard:
  - Dublin Core Metadata Initiative
- The language resources community can simply augment the generic standard
  - OLAC has done this
OLAC metadata standard

- Dublin Core metadata standard has:
  - 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 (ISO 639-3), Linguistic Data Type, Linguistic Field, Participant Role, Discourse Type
Submission protocol

3. Institutions with language resources to share need an open protocol for submitting metadata to the aggregator.

- The digital library community has developed one:
  - Open Archives Initiative (OAI) protocol

- Our community can simply adapt it
  - OLAC has done this
  - Specifies an exact syntax for resource description
A metadata record as submitted

- <olac:olac xsi:schemaLocation="http://www.language-archives.org/OLAC/1.0/
  http://www.language-archives.org/OLAC/1.0/olac.xsd
  http://purl.org/dc/elements/1.1/
  http://www.language-archives.org/OLAC/1.0/dc.xsd http://purl.org/dc/terms/
  http://www.language-archives.org/OLAC/1.0/dcterms.xsd">
  <title>Ega lexicon (Gbery)</title>
  <creator>Gbery, Eddy Aime</creator>
  <creator>Baze, Lucien</creator>
  <subject xsi:type="olac:language" olac:code="ega"/>
  <description>Ega lexicon in Shoebox format</description>
  <publisher>unpublished</publisher>
  <contributor>Lindenlaub, Juliane</contributor>
  <date>2003-03</date>
  <type xsi:type="olac:linguistic-type" olac:code="lexicon"/>
  <format>shoebox</format>
  <language xsi:type="olac:language" olac:code="fra"/>
  <language xsi:type="olac:language" olac:code="ega"/>
  <language xsi:type="olac:language" olac:code="eng"/>
  <language xsi:type="olac:language" olac:code="deu"/>
  <coverage>Cote d'Ivoire</coverage>
</olac:olac>

14 July 2007
Harvesting protocol

4. Institutions who want to provide filtering services need an open protocol for harvesting metadata from the aggregator.

- The digital library community has developed one:
  - OAI Protocol for Metadata Harvesting
- Our community can simply adopt this
  - OLAC has done this
  - Two institutions now provide search over the 30,000 aggregated resources
    - Linguist List and Linguistic Data Consortium
OAI Protocol for Metadata Harvesting

- There are six verbs:
  - GetRecord, Identify, ListIdentifiers, ListMetadataFormats, ListRecords, ListSets

- Requests expressed as URLs:
  - `baseURL?verb=value&parameters`

- For instance:

- Answer returned as an XML document
A metadata record as displayed

Document Information

General Description:

Title: Ega lexicon (Gbery)

Archive: U Bielefeld Language Archive

Archive URL: http://www.spectrum.uni-bielefeld.de/langdoc/

Creator(s): Gbery, Eddy Alme
            Baze, Lucien

Description: Ega lexicon in Shoebox format

Contributor(s): Lindenlaub, Juliane

Date: 2003-03

Coverage: Cote d'Ivoire

Format: shoebox

Language: French [fra]
          Ega [ega]
          English [eng]

14 July 2007
A listing of search results

Search
OLAC: potawatomi Find -- All archives --

Search results for "potawatomi" in all OLAC archives 9 results from 4 archive(s)

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)

1. ★ oai:perseus.tufts.edu:Perseus:text:2000.03.0068 Similar records by: score language type
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 Winnipeck, 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.
Basic OLAC infrastructure
Subcommunities

7. Subcommunities need to establish conventions within the metadata standard in order to support specialized filtering, gateways, and spiders.

■ Subcommunities have more specialized filtering needs than the OLAC standard supports
  ► Just as OLAC was immediately successful because it specialized existing standards (OAI, DC)
  ► Subcommunities can do the same by specializing the OLAC standard
  ► Then build services that exploit the specializations
Examples of subcommunities

- LL-MAP project using extensions to the Coverage element for geospatial coordinates
  - E.g., `<coverage xsi:type="dcmi:point"> name=Perth, W.A.; east=115.85717; north=-31.95301 </coverage>`

- This workshop: *Applying Cyber-infrastructure to the Language Sciences: A Case Study for Language Acquisition*

- Can be defined by a fixed metadata value
  - E.g., `<format>LIFT 1.0</format>`
Datum-level search

8. Subcommunities need to establish content standards in order to support datum-level search over relevant resources harvested from the aggregator.

- We also want to search across resources for matching data within the content.
- Such microinteroperation implies that the subcommunity interested in a particular content type must establish standards for markup of resources (including controlled vocabularies).
Implementing datum-level search

- Aggregate and filter at a more granular level:
  1. Query aggregator for all subcommunity resources
  2. Harvest the resources themselves
  3. Parse content into a content-specific database
  4. Implement search service over that database
  5. Submit descriptions of the reports generated
Cyberinfrastructure for linguistics
Next steps

■ Infrastructure for macrointeroperation
  ► How will we identify subcommunity resources?
    ▪ Reserved values for <format>, <type> or <subject>
    ▪ A formal extension vocabulary for any of these
  ► Who will implement a subcommunity search?

■ Infrastructure for microinteroperation
  ► What standards for content are agreed on?
    ▪ Markup schemas
    ▪ Controlled vocabularies
  ► Who will implement services that exploit these?

■ The OLAC administrators stand by to advise