The Seven Pillars of Open Language Archiving: A Vision Statement

Gary Simons and Steven Bird

Workshop on Web-based Language Documentation and Description
12 December 2000

Overview: a gap analysis

- What users want—the ideal
- What users actually get —the gap
- What it would take to bridge the gap—a proposed community infrastructure

A building metaphor

The infrastructure is erected on seven pillars:

- Data
- Tools
- Advice
- Gateway
- Metadata
- Review
- Standards

What users want

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

- Data
- Tools
- Advice

1. Data

- Information that documents or describes a language of interest
- A wide variety of formats: print publications, computer data files, sound recordings, hand-written index cards, and so on
- A wide variety of content: word lists, paradigms, texts, annotations, lexicons, grammar descriptions, and so on

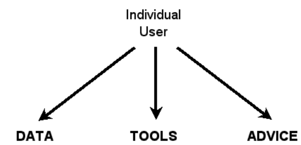
2. Tools

- Computational resources that facilitate creating, viewing, querying, or otherwise using language data
- These include: application programs, components, fonts, style sheets, document type definitions, and so on

3. Advice

- Information that users would typically solicit when they need help
- For instance,
  - What data sources should I rely on?
  - What software tools should I use?
  - What practices should I follow when creating data? When using data?

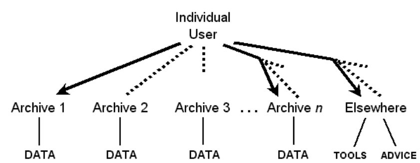
The ideal situation



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 hundreds of other sites

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 that provides four things:

- A single gateway
- Uniform metadata
- A review process
- Standards

4. Gateway

- A single portal through which users gain access to all available data, tools, and advice
- The actual data, tools, and advice are located on hundreds of sites all over the Internet—the gateway stores links to them.
- By accessing the single gateway site, the user gains access to all available data, tools, and advice.

5. Metadata

- Uniform descriptions of all available data, tools, and advice
- Not the data itself, but data about the data; thus it works for digital and non-digital holdings.
- Use specialized metadata elements to meet requirements specific to language archives:
  - uniformly identifying languages
  - matching data formats to the appropriate tools

6. Review

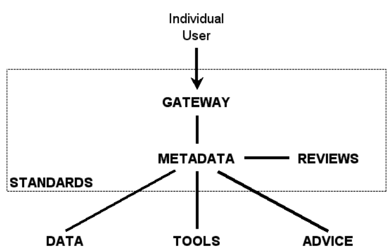
- Peer evaluation of available data, tools, and advice
- Peer review is an important function of any academic community.
- Review by individuals (with responses)
- Review by community to establish some advice as recommended best practice

7. Standards

The framework that allows the core infrastructure to function:

- Gateway—governed by a protocol for harvesting metadata from participating archives
- Metadata—governed by an XML schema that ensures uniformity across all archives
- Review—governed by a process that promotes draft to candidate and then to best practice

Proposed community infrastructure



Open Language Archives Community

A community of on-line archives and services that meet the needs of the language documentation community by:

- Implementing the Open Archives Initiative protocol for harvesting metadata
- Maintaining a community-specific metadata set
- Maintaining a review process for the community

Going deeper

- A mockup of the OLAC gateway
- The Open Archives Initiative
- Prototypes based on OAI
- A trio of documents:
  - Requirements
  - Survey
  - White paper