### The OLAC Metadata Set and Controlled Vocabularies

**Steven Bird**  
Penn  

**Gary Simons**  
SIL

### Types of Language Resource

**DATA:** any information which documents or describes a language, such as a:
- monograph, data file, shoebox of index cards, unanalyzed recordings, heavily annotated texts, complete descriptive grammar

**TOOLS:** computational resources that facilitate creating, viewing, querying, or otherwise using language data
- includes fonts, stylesheets, DTDs, Schemas

**ADVICE:** any information about:
- reliable data sources, appropriate tools and practices

### The Language Resources Community

**Creators and Users of Language Resources:**
- speakers, educators, linguists, technologists

**Immediate Infrastructure:**
- archivists, software developers, publishers

**Sponsors & Promoters:**
- professional associations, funding agencies, non-governmental organizations

**Scale:** tens of thousands of people

### Now: Underdevelopment

- **The building blocks**
  - data, formats, tools, interfaces
  - diversity & incompatibility
  - the pieces fit together poorly

- **Resource discovery**
  - "word of mouth" (e.g. CORPORA)
  - search engines
  - low precision and recall

- **Architecture**
  - small, unstable, unscalable
  - exchange and reuse of "primary materials"
  - diversity is restricted
Future: Development

- The building blocks
  - data, formats, tools, interfaces
  - diversity with compatibility
  - the pieces fit together well
- Resource discovery
  - resources in federated archives
  - common finding aids
  - high precision and recall
- Architecture
  - large, stable, scalable
  - aggregation and integration of complex structures and services
  - diversity is facilitated

Three Approaches to Bridging the Gap

1. Monolithic ✽
2. Independent ✽
3. Coordinated ✫

The Gap

Monolithic Approach

"One day, a single, massive project will succeed in bridging the gap"

Analogy: a centralized database as a complete information system
Independent Approach

"Given enough time, the accretion of independent initiatives will bridge the gap"

Analogy: the world-wide web as a complete information system

Coordinated Approach

"A shared architectural vision, having many components, and implemented in stages by the community, will bridge the gap"

Analogy: federated databases; semantic web

The Foundation: 3 initiatives

1. **Dublin Core Metadata Initiative (DC)**
   - founded in 1995 (Dublin, Ohio)
   - conventions for resource discovery on the web

2. **Open Archives Initiative (OAI)**
   - founded in 1999 (Santa Fe)
   - interoperability of e-print services

3. **Open Language Archives Community (OLAC)**
   - founded in 2000 (Philadelphia)
   - a partnership of institutions and individuals
   - creating a worldwide virtual library of language resources

Foundation 1: DC Elements

15 metadata elements:
- broad interdisciplinary consensus
- each element is optional and repeatable
- applies to digital and traditional formats
- Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage, Rights.

dublincore.org
Foundation 1: DC Qualifiers

Encoding Schemes:
- a controlled vocabulary or notation used to express the value of an element
- helps a client system to interpret the element content
- e.g. Language = "en" (not "English", "Anglais", ...)

Refinements:
- makes the meaning of an element more specific
- e.g. Subject.language, Type.linguistic

Foundation 2: OAI Standards

To implement the OAI infrastructure, an archive must comply with two standards:

1. The OAI Shared Metadata Set
   - Dublin Core
   - interoperability across all repositories

2. The OAI Metadata Harvesting Protocol
   - HTTP requests - 6 verbs:
     - Identify, ListIdentifiers, ListMetadataFormats, ListSets, ListRecords, GetRecord
   - XML responses

Foundation 2: OAI Repository

<table>
<thead>
<tr>
<th>ITEM</th>
<th>Unqiue Identifier</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Metadata record (DC)</td>
</tr>
<tr>
<td></td>
<td>Metadata record (other format)</td>
</tr>
</tbody>
</table>

Describes

ARCHIVE

<table>
<thead>
<tr>
<th>HOLDING</th>
<th>Document</th>
<th>Recording</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Data</td>
<td>Artifact</td>
</tr>
<tr>
<td></td>
<td>Software</td>
<td>Surrogate</td>
</tr>
</tbody>
</table>

Foundation 2: OAI Service Providers and Data Providers

Users

Service Provider

Data Provider
Foundation 3: OLAC

OLAC was founded at the Workshop on Web-Based Language Documentation and Description (Philadelphia, 2000)

- sponsored by NSF: TalkBank, ISLE, IRCS
- 100 participants:
  - computational linguists, descriptive linguists, archivists
  - N America, S America, Europe, Africa, Middle East, Asia, Australia

Aside: OLAC Organization

- **Coordinators:** Steven Bird & Gary Simons
- **Advisory Board:** Helen Arista Dry, Susan Hockey, Chu-Ren Huang, Mark Liberman, Brian MacWhinney, Michael Nelson, Nicholas Ostler, Henry Thompson, Hans Uszkoreit, Antonio Zampolli
- **Participating Archives & Services:** LDC, ELRA, DFKI, CBOLD, ANLC, LACITO, Perseus, SIL, APS, Utrecht
- **Prospective Participants:** ASED A, Academia Sinica, AISRI, INALF, LCAAJ, Linguist, MPI, NAA, OTA, Rosetta, Tibetan Digital Library
- **Working Groups:** 5 set up at Philadelphia workshop - but focus has been on infrastructure and metadata
- **Individual Members:** ~120

Foundation 3: OLAC Aims

OLAC, the Open Language Archives Community, 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.

Foundation 3: OLAC & OAI

Recall: OAI data providers must support:

- Dublin Core Metadata
- OAI Metadata harvesting protocol

BUT: OAI data providers can support:

- a more specialized metadata format
- a more specialized harvesting protocol

What OLAC does:

- specialized metadata for language resources
- specialized harvesting (extra validation)
The OLAC Metadata Set

The three categories of metadata:
- **Work language**: describes information entities and their intellectual attributes
  - e.g. names of works and their creators
- **Document language**: describes and provides access to the physical manifestation of information
  - e.g. format, publisher, date, rights
- **Subject language**: describes what a document is about
  - e.g. subject, description


OLACMS Work Language

e.g. Creator:
- Def: An entity primarily responsible for making the content of the resource
- Text to name the creator
  - e.g. BCP: "Surname, Firstname"
- Refinement to Dublin Core: OLAC-Role
- OLAC-Role is a controlled vocabulary
  - *author, editor, translator, transcriber, sponsor, ...*
OLACMS Document Language

e.g. Format markup:
- Def: The OAI identifier for the definition of the markup format
- references the DTD, Schema, or some other definition of the markup format
  - e.g. oai:nist:timit86
- For software: supported markup formats
- Consequences:
  - Ensures that format definitions are archived
  - Queries can do a join to find data of a given type for which software is available

OLACMS: Subject Language

e.g. Subject language
- Def: A language which the content of the resource describes or discusses
- Starting points:
  - ISO 639, LANGIDs, RFC-3066 (1766), Ethnologue
  - Unicode Consortium & IETF
  - aware of shortcomings of RFC-3066
  - want to incorporate Ethnologue codes
- Current proposal being considered
  - 4-letter codes (Ethnologue 3-letter codes plus prefix)
  - where an unambiguous 2 or 3-letter code exists, use it, and drop the Ethnologue equivalent
- Other developments:
  - LINGUIST Ancient Languages: x-ll-xakk = Akkadian
  - UCSB workshop discussed Language Code Consortium

OLACMS: Subject Language

E.g. Type lingdata (was type data)
- Def: The nature or genre of the content of the resource, from a linguistic standpoint.
- Difficult: "CL does not yet have a systematics or classification scheme" (Uszkoreit)
- Encoding scheme: OLAC-LingData (OLAC-Data)
- Primary classification:
  - transcription: a time-ordered symbolic representation of a linguistic event
  - annotation: any kind of structured linguistic information that is explicitly aligned to some spatial and/or temporal extent of a linguistic record
  - description: any description or analysis of a language (structure is independent of the linguistic events)
  - lexicon: any record-structured inventory of forms

OLACMS: Subject Language

E.g. Secondary classification for transcription
- transcription/orthographic
- transcription/phonetic
- transcription/prosodic
- transcription/morphological
- transcription/gestural
- transcription/part-of-speech
- transcription/syntactic
- transcription/discourse
- transcription/musical
OLAC MHP 1:
Representing the Metadata

See Figure 5 in the proceedings paper

Refinements:
  <Creator refine="Author">Bateman, John</Creator>

Encoding scheme:
  <Format.os code="Unix/Solaris"/>

Language:
  <Description lang="fr">Une description de la resource ecrit en Francais</Description>

Header:
  xmlns="http://www.language-archives.org/OLAC/0.3/"

OLAC MHP 2:
Refinements to OAI Protocol

1. Identify
   • specify the format of the archive self-description field

2. ListMetadataFormats
   • specify that OLAC is one of the returned formats and that the URL points to the canonical schema

3. ListIdentifiers
   • when OLAC is specified as the required metadata format, ensure that the repository returns at least one record identifier

OLAC Process

Lays out the core values of OLAC:
  • openness, consensus, empowering the players, peer review

Describes the organization of OLAC:
  • coordinators, advisory board, participating archives and services, prospective participants, working groups, participating individuals

Defines processes for documents and working groups

http://www.language-archives.org/OLAC/process.html

Summary: Three Standards Define the Community
Third Layer: OLAC BCPs

Recommendations for appropriate use

1. OLAC Metadata Set:
   - e.g. don't abbreviate association names:
     - `<publisher>Association for Computational Linguistics</publisher>`

2. OLAC MHP:
   - e.g. where possible map a language designation to a code in OLAC-Language, instead of freeform text

3. OLAC Process:
   - e.g. use such-and-such an XML format for archiving wordnets

Fourth Layer: Software

Beginning with any kind of language resource, there will be software to:

- convert it to archival format (if possible)
  - e.g. replace legacy fonts with Unicode
- create a metadata record
  - e.g. LDC's metadata lives in an Oracle database
- export this record to XML
  - "publish" the record in the OLAC format
- harvest the record
  - service provider software to retrieve the record and present it to end-users

Summary: Standards are Supplemented with Community Favoured Syntax and Semantics

Summary: With the software in place, we have a complete platform
Summary: Repositories completely bridge the gap, letting us consistently organize and archive our resources

Sixth Layer: OLAC Services

1. Metadata Validation
   - a public interface which permits humans and machines to verify that a putative OLAC record is valid
2. Registration Server
   - tests for OAI membership
   - tests conformance with the MHP:
     - responses to verbs, metadata validation
   - creates a record for the repository: service providers can discover what repositories exist
3. Archive Summarization
   - archive self-description, statistics

Seventh Layer: User Services

1. Union Catalog
   - a single place to query all participating archives
   - LINGUIST will host the primary service provider, guaranteed to be complete
2. Peer Review
   - all archive records and holdings will be open for signed peer review
   - will provide community recognition for resource creation work
3. Interface for metadata submission
   - a proliferation of small repositories
   - create some XML and submit the URL

Summary: Seven Layers Complete the Bridge

- Repositories
- OLAC REPOSITORIES
- OLAC CONTENT
- OLAC CREATE
- OLAC METADATA
- OLAC EXPORT
- OLAC FORMAT
- OLAC DELIVER

- OLAC SERVICES
- OLAC USER SERVICES

- Initiatives
- Standards
- Recommendations
- Software
Potential Criticisms 1

Aren't you converting the bazaar into a cathedral?
- it wasn't a bazaar - there were no universal currencies or languages
- it won't be a cathedral - the result will be more diverse than what we began with

Moving Forward...

The Coordinated Approach:
"A shared architectural vision, having many components, and implemented in stages by the community, will bridge the gap"

Do you share this vision?
- NO: what do we need to discuss or change?
- YES: how do you want to participate?
  - set up a repository (join OLAC-Implementers)
  - sign up as an individual (join OLAC-General)
  - help set up the controlled vocabularies (join or create a working group)

Potential Criticisms 2

There's too much infrastructure here - it will be impossible to get started!
- Metadata elements are all optional
- The MHP is lightweight (CGI + simple XML)
- open source implementations are available (Perl, PHP, Java, XSLT)
- OLAC already has 10 participating repositories (i.e. we've prototyped many parts of the bridge)

OLAC

Acknowledgements: ISLE and TalkBank projects (NSF), participants of the Philadelphia workshop, Eva Banik (programmer), Hernando de Soto (the analogy)