White Paper on Establishing an Infrastructure for Open Language Archiving

Steven Bird and Gary Simons

The Open Archives Initiative

www.openarchives.org

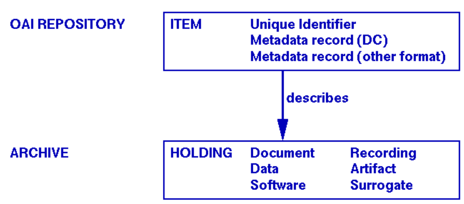
- Began with e-prints
- Now covers digital repositories of scholarly materials, regardless of type

Each participating archive implements a repository:

- Item: identifier + metadata
- Specifies entry point

WP:1

OAI Repositories and Archives



WP:1

Built on Two Standards

The OAI shared metadata set: Dublin Core

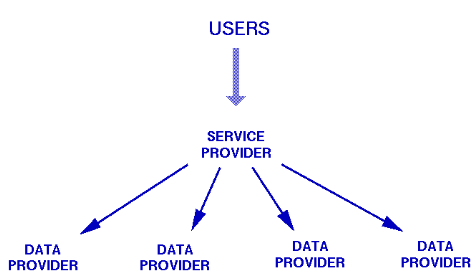
- core set of 15 metadata elements
- represent a broad, interdisciplinary consensus
- widely useful for resource discovery

OAI Metadata Harvesting Protocol

- software services can query a repository
- retrieve item identifiers and metadata records

WP:1

OAI Service and Data Providers



WP:1

Definition of the OAI Community

The OAI is a community of archives which:

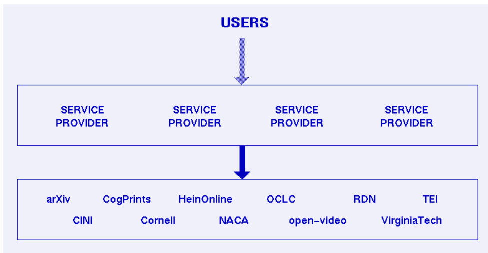
- supply Dublin Core metadata
- support the OAI Metadata Harvesting Protocol
- register with the OAI

Any compliant repository can register

No other notion of community membership

WP:2

The OAI Community



WP:2

OAI Supports Specialist Communities

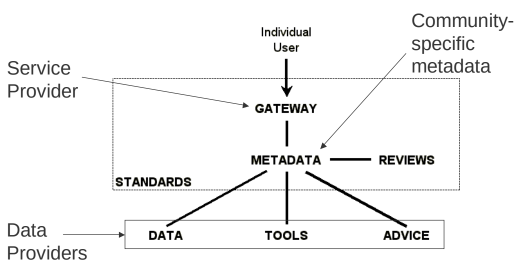
The community can define metadata formats other than Dublin Core

- Specific to a particular domain
- DPs serve the new format
- SPs harvest the new format

Result: an *OAI subcommunity*

WP:2

What does OAI provide us?



WP:2

Proposed OLAC Metadata Set

Metadata is what makes OLAC a distinct subcommunity of the OAI

- Through metadata, our community describes the resources which are fundamental to the enterprise of language documentation
- Minimally extend Dublin Core to express what is fundamental about:
  - Open
  - Language
  - Archiving
- But how?

WP:3

Back to the Requirements

4. Identify the languages that archived items relate to
5. Identify how open or restricted an item is
6. Identify format and encoding details for digital resources
7. Identify other resources required for using an item
8. Match data resources with appropriate software tools

WP:3.2

OLAC metadata elements

- Subject.language
- Rights.openness
- Format.openness
- Format.encoding
- Format.markup
- Type.data
- Relation.requires
- Rights.openness
- Format.language
- Type.functionality
- Type.os
- Type.osversion
- Type.cpu

WP:3.2-3

Controlled vocabulary servers

Many elements have a restricted range of values:

- Rights.openness: *open, published, restricted, unknown*
- Subject.language: *6000+ Ethnologue codes*

Controlled vocabulary server:

- Network-accessible service
- Maintains and documents a vocabulary
- SIL has agreed to be a C.V.S. for language id

WP:3.5

Subcommunities with richer metadata standards

Just as OLAC is a subcommunity of the OAI, there are other subcommunities in the scope of OLAC

- Language data centers (LDC, ELRA, GSK)
- ISLE Meta Data Initiative – *detailed metadata for describing recorded speech events*

These subcommunities would support DC and OLAC metadata, plus their own set

- Specialized service provider
- Focussed searching based on richer metadata

WP:3.6

Founding the Open Language Archives Community

- Standards
- OLAC definition
- OLAC Gateway
- Primary OLAC service provider
- Peer review
- Defining recommended best practice

WP:4

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

WP:4.1

OLAC Definition

Definition: *The Open Language Archives Community (OLAC) is the community of language archives and associated services which implement the OLAC standards.*

Purpose: *to support the language documentation community, by fostering the sharing of language resources.*

Advisory council: *each OLAC archive will be asked to select a representative to serve on an advisory council.*

WP:4.2

OLAC Gateway: www.language-archives.org

This site will host information for the community of people:

- OLAC standards documents
- index of service providers
- collection of best practice recommendations

…plus information for the community of machines:

- OLAC metadata schema
- registry of data providers
- controlled vocabulary servers (local or remote)

WP:4.3

Primary OLAC Service Provider

THE LINGUIST LIST
EASTERN MICHIGAN UNIVERSITY ■ WAYNE STATE UNIVERSITY

Qualifications:
- foremost electronic network of linguists, with over 13,000 members worldwide
- a decade of experience
- worldwide mirrors

Roles:
- Provides a complete union catalog
- Institutes an informal, open, peer-review process

WP:4.4

Peer Review

How can you judge the quality of a digital resource?
- scale, quality, openness of the resource / support
- information may be misleading, outdated, erroneous
- access delayed/blocked by unadvertised restrictions
- problems with data, tools, formats, best practices

An informal, open, peer review process
- Users of a data or service provider can report their experience using a form on the OLAC Gateway
- Review forwarded to the provider, post a response
- Visitors to the Gateway could peruse them

WP:4.5

Defining recommended best practice

Anyone could submit an RFC, posted on Gateway
- RFC: existing practice; experience; case for wider adoption
- RFCs would be reviewed by the community and the advisory council
- Accepted RFCs promoted to the status of "Recommended Best Practice"

Not standards, but recommendations
- To limit the needless incompatibilities of format
- Encourage genuine innovation

WP:4.6

Next steps: This week

- Working group discussions, leading to revised requirements
- Working group discussions, leading to a revised white paper
- Identify alpha test group
- Endorsement and announcement

WP:5