OLAC: The Open Language Archives Community

Steven Bird
Penn

Gary Simons
SIL

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

Reading

Bird & Simons (2001) The OLAC metadata set and controlled vocabularies. *ACL Workshop on sharing tools and resources for research and education.* http://arXiv.org/abs/cs/0105030

www.language-archives.org

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

Metadata: Necessary?

The goals: finding, collocating, choice, acquisition, navigation

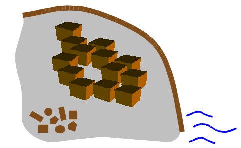
Against:

- cost, user's ability to exploit the metadata, not needed for some purposes

For:

- comprehensive retrieval (collocation) e.g. historian, mathematician, inventor
- user's abilities are generally poor (choice of search terms, refining the search)

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*

Metadata: Cost Issue

Technical solution:

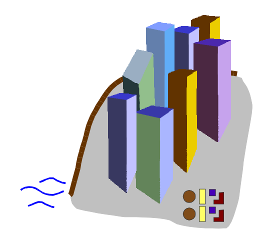
- automatic extraction of metadata
- mitigate the costs

Political solution:

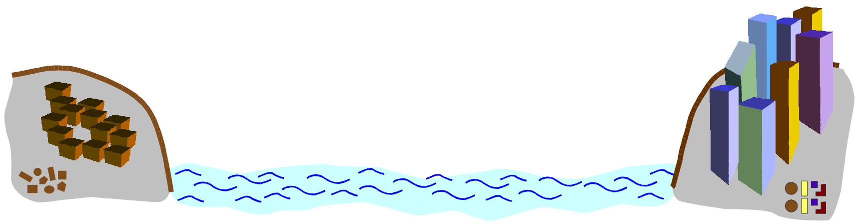
- standards in support of cooperative efforts
- distribute the costs

Future: Development

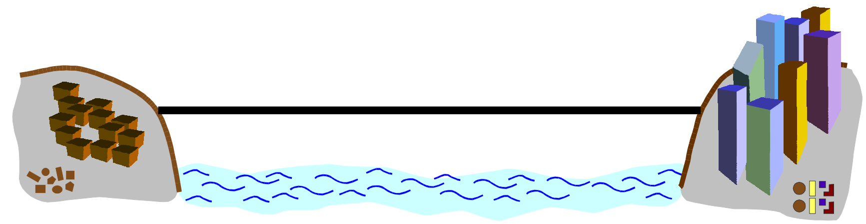
- The building blocks
  - data, formats, tools, interfaces
  - diversity with compatibility
  - *the pieces fit together well*
- Resource discovery
  - resources in federated archives
  - common finding aids
  - *high precision and recall*
- Architecture
  - large, stable, scalable
  - aggregation and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



The Gap



Monolithic Approach



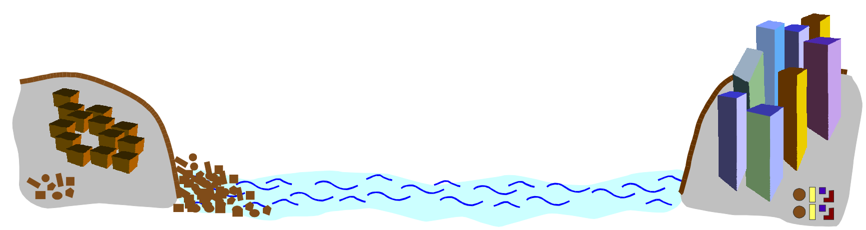
"One day, a single, massive project will succeed in bridging the gap"

Analogy: a centralized database as a complete information system

Three Approaches to Bridging the Gap

1. **Monolithic** ☆
2. **Independent** ☆
3. **Coordinated** ✚

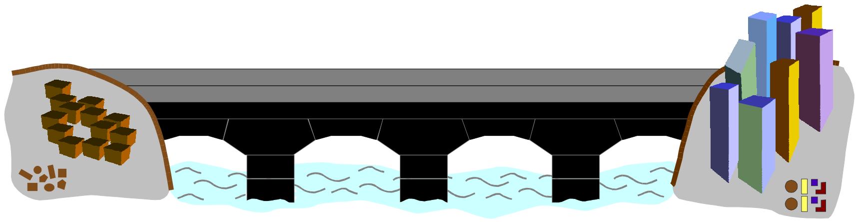
Independent Approach



"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"

Analogies: federated databases; semantic web

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

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

DC: Title Element

Title: A name given to the resource.

Comments: Typically, a Title will be a name by which the resource is formally known.

Example:

<title>A Dictionary of the Nggela Language</title>

DC: Creator Element

Creator: An entity primarily responsible for making the content of the resource.

Comments: Examples of a Creator include a person, an organization, or a service.

Example:

<creator>Bloomfield, Leonard</creator>

DC: Description Element

Description: An account of the content of the resource.

Comments: Description may include an abstract, table of contents, reference to a graphical representation of the content, or a free-text account.

Example:

<description>The CALLHOME Japanese corpus of telephone speech consists of 120 unscripted telephone conversations between native speakers of Japanese. …</description>

DC: Subject Element

Subject: The topic of the content of the resource.

Comments: Typically, a Subject will be expressed as keywords, key phrases or classification codes.

Example:

<subject>Czech</subject>

DC: Publisher Element

Publisher: An entity responsible for making the resource available.

Comments: Examples of a Publisher include a person, an organization, or a service.

Example:

<publisher>Oxford University Press</publisher>

DC: Contributor Element

Contributor: An entity responsible for making contributions to the content of the resource.

Comments: Examples of a Contributor include a person, an organization, or a service.

Refinements: author, editor, translator, transcriber, sponsor, …

Example:

<contributor refine="funder">National Science Foundation</contributor>

DC: Type Element

Type: The nature or genre of the content of the resource.

Comments: Type includes terms describing general categories, functions, genres, or aggregation levels for content. (Distinct from physical manifestation.)

Example:

<type>image</type>

DC: Date Element

Date: A date associated with an event in the life cycle of the resource

Comments: Use the YYYY-MM-DD format defined by the W3C Date-Time Format

Example:

<date>1996-10-16</date>

DC: Format Element

Format: The physical or digital manifestation of the resource.

Comments: Typically, Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware, or other equipment needed to display or operate the resource.

Example:

<format>5,237 entries in a 1.2Mb XML file</format>

DC: Identifier Element

Identifier: An unambiguous reference to the resource within a given context.

Comments: Formal identification systems include URI, DOI, ISBN. For conventional archives, identifier may give a local shelf or box number.

Example:

<identifier>http://arXiv.org/abs/cs.CL/0010033</identifier>

DC: Language Element

Language: A language of the intellectual content of the resource.

Comments: Language is used for a language the resource is in, as opposed to the language it describes. The creator of the resource assumes that users will understand this language.

Example:

<language>Czech</language>

DC: Source Element

Source: A reference to a resource from which the present resource is derived.

Comments: This is for a "derivative work", which is a transformation of the source work, e.g. by translation, abridgement, dramatization, recording, transcription, digital encoding, editorial revision, annotation, elaboration, etc.

Example:

<source>oai:somearchive:holding123</source>

DC: Relation Element

Relation: A reference to a related resource.

Comments: Relation documents relationships between resources, e.g. aggregation, required software/data.

Refinements: IsVersionOf, HasVersion, IsReplacedBy, Replaces, IsRequiredBy, Requires, IsPartOf, HasPart, IsReferencedBy, References, IsFormatOf, HasFormat

Example:

<Relation refine="Requires">CommonLisp</Relation>

DC: Coverage Element

Coverage: The extent or scope of the content of the resource.

Comments: Coverage typically includes spatial location, temporal period, or jurisdiction.

Example:

<coverage>New England</coverage>

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

DC: Rights Element

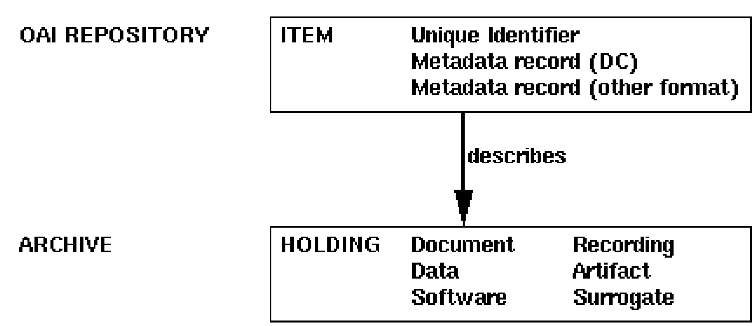
Rights: Information about rights held in and over the resource.

Comments: This is a rights management statement for the resource, or a reference to a service providing such information. It may cover Copyright, IPR, and other property rights.

Example:

<rights>Copyright (C) 2001 Steven Bird, distributed under OPL</rights>

Foundation 2: OAI Repository



Foundation 2: OAI Standards

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

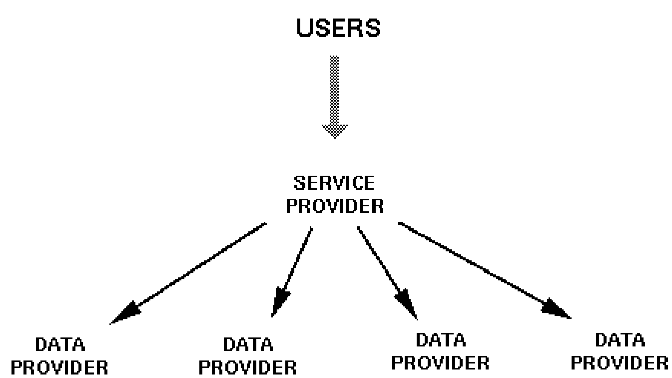
1. The OAI Shared Metadata Set
   - Dublin Core
   - interoperability across all repositories
2. The OAI Metadata Harvesting Protocol
   - HTTP requests - 6 verbs:
     - Identify, ListIdentifiers, ListMetadataFormats, ListSets, ListRecords, GetRecord
   - XML responses
   - Demonstration

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

Foundation 2: OAI Service Providers and Data Providers



Aside: OLAC Organization

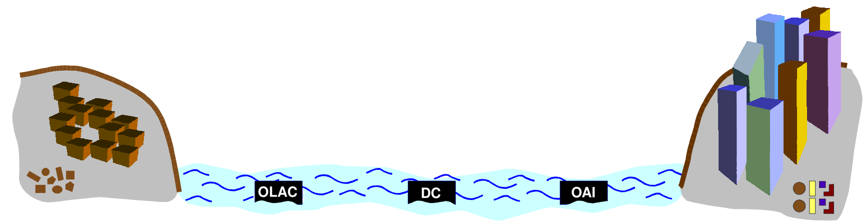
- **Coordinators:** Steven Bird & Gary Simons
- **Advisory Board:** Helen Aristar Dry, Susan Hockey, Chu-Ren Huang, Mark Liberman, Brian MacWhinney, Michael Nelson, Nicholas Ostler, Henry Thompson, Hans Uszkoreit, Antonio Zampolli
- **Participating Archives & Services:** LDC, ELRA, DFKI, CBOLD, ANLC, LACITO, Perseus, SIL, APS, Utrecht
- **Prospective Participants:** ASEDA, 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.

Summary: Three Initiatives Provide the Foundation



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)

Next Layer: OLAC Standards

Aside:

- standards = the protocols and interfaces that allow the community to function
- recommendations = "standards" for representing linguistic content

OLAC has three primary standards:

- *OLACMS:* the OLAC Metadata Set (Qualified DC)
- *OLAC MHP:* refinements to the OAI protocol
- *OLAC Process:* a procedure for identifying Best Common Practice Recommendations

The OLAC Metadata Set

The three categories of metadata:

- Work language: describes information entitites 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

cf: Svenonius (2000) The Intellectual Foundation of Information Organization (MIT Press)

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 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: Subject Language

E.g. Type.lingdata (was type.data)

- Def: The nature or genre of the content of the resource, from a linguistic standpoint.
- 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/"

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*

OLAC MHP 2: Refinements to OAI Protocol

1. Identify

- specify the format of the archive self-description field

2. ListMetadataFormats

- specify tha 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

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>

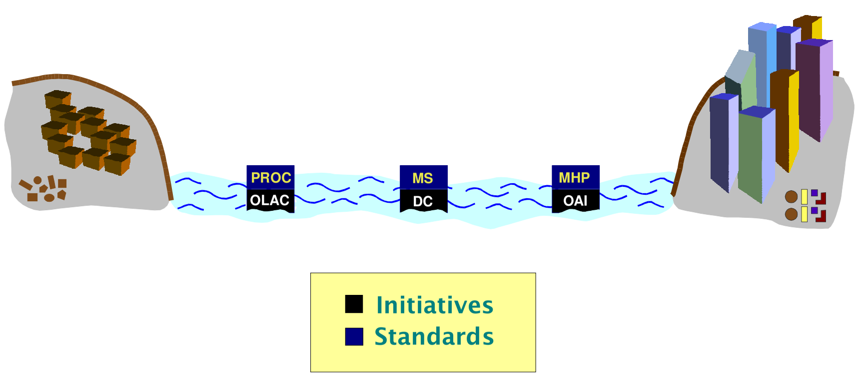
2. OLAC MHP:

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

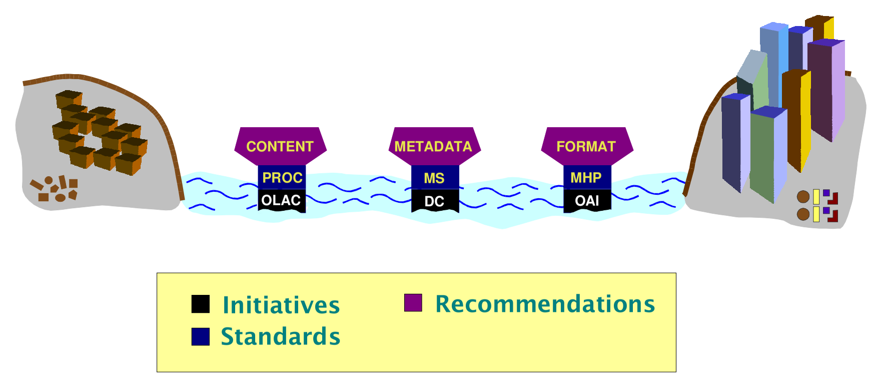
3. OLAC Process:

- e.g. use such-and-such an XML format for archiving wordnets

Summary: Three Standards Define the Community



Summary: Standards are Supplemented with Community Favoured Syntax and Semantics

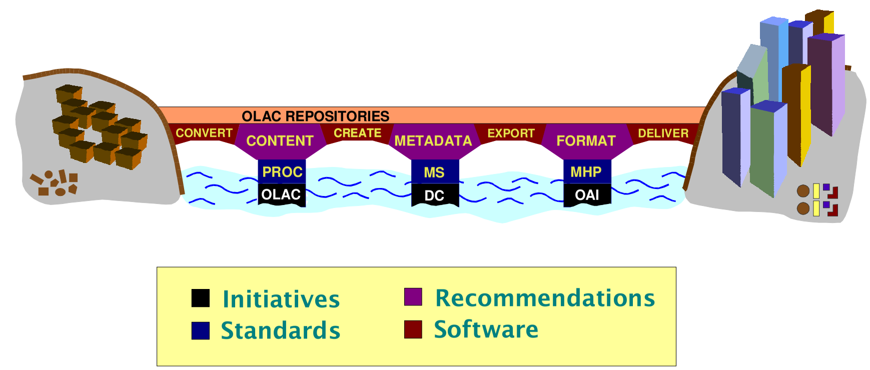


Fourth Layer: Software

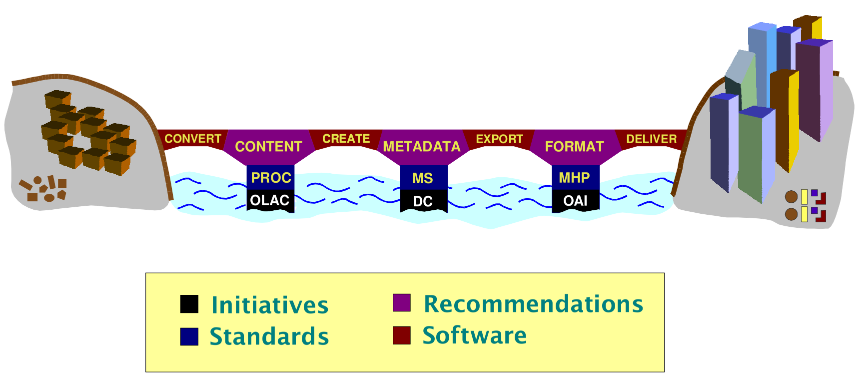
Beginning with any kind of language resource, there will be software to:

- 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: Repositories completely bridge the gap, letting us consistently organize and archive our resources



Summary: With the software in place, we have a complete platform



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

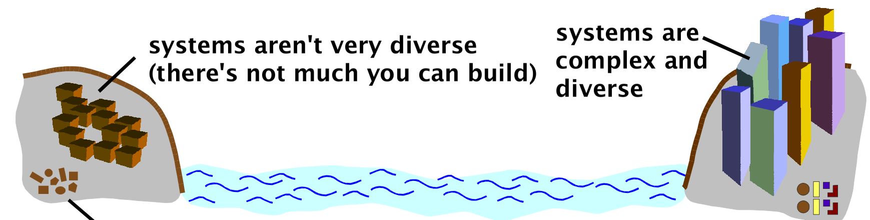
Seventh Layer: User Services

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

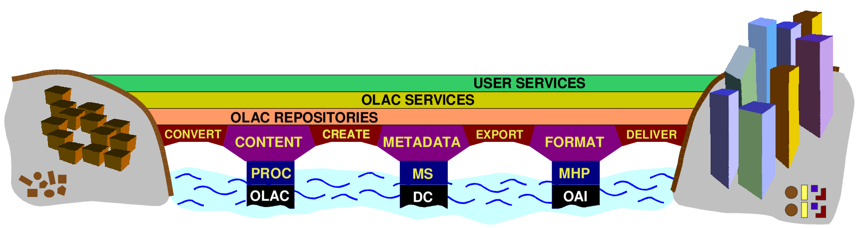
Potential Criticisms 1

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



Summary: Seven Layers Complete the Bridge



Initiatives	Recommendations
Standards	Software

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)

Demonstration

Moving Forward...

The Coordinated Approach:

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

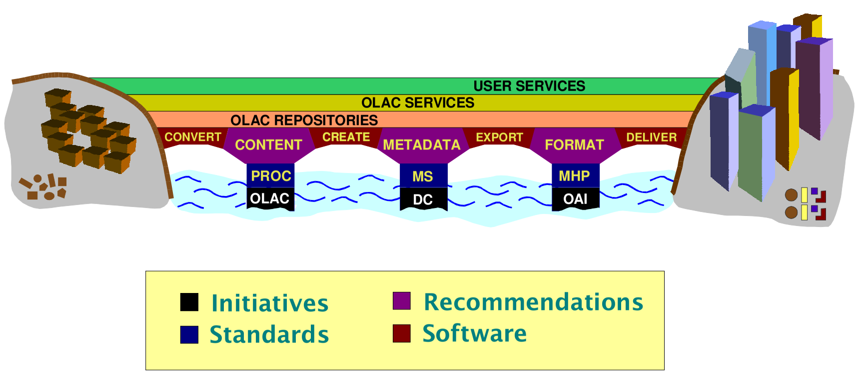
Do you share this vision?

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)

OLAC



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