The Open Language Archives Community and Asian Language Resources

Steven Bird
Linguistic Data Consortium
University of Pennsylvania
3615 Market Street
Philadelphia, PA 19104, USA
sb@ldc.upenn.edu

Gary Simons
SIL International
7500 West Camp Wisdom Road
Dallas, TX 75236, USA
Gary_Simons@sil.org

Chu-Ren Huang
Institute of Linguistics
Academia Sinica
115 Nankang, Taipei, Taiwan
churen@gate.sinica.edu.tw

Abstract

The Open Language Archives Community (OLAC) is a new project to build a worldwide system of federated language archives based on the Open Archives Initiative and the Dublin Core Metadata Initiative. This paper aims to disseminate the OLAC vision to the language resources community in Asia, and to show language technologists and linguists how they can document their tools and data in such a way that others can easily discover them. We describe OLAC and the OLAC Metadata Set, then discuss two key issues in the Asian context: language classification and multilingual resource classification.

1 Introduction

Language technology and the linguistic sciences are confronted with a vast array of *language resources*, richly structured, large and diverse. Multiple *communities* depend on language resources, including linguists, engineers, teachers and actual speakers. Many individuals and institutions provide key pieces of the infrastructure, including archivists, software developers, and publishers. Today we have unprecedented opportunities to *connect* these communities to the language resources they need.

We can observe that the individuals who use and create language resources are looking for three things: data, tools, and advice. By *data* we mean any information that documents or describes a language, such as a published monograph, a computer data file, or even a shoebox full of hand-written index cards. The information could range in content from unanalyzed sound recordings to fully transcribed and annotated texts to a complete descriptive grammar. By *tools* we mean computational resources that facilitate creating, viewing, querying, or otherwise using language data. Tools include not just software programs, but also the digital resources that the programs depend on, such as fonts, stylesheets, and document type definitions. By *advice* we mean any information about what data sources are reliable, what tools are appropriate in a given situation, what practices to follow when creating new data, and so forth (e.g. the Corpora List archives [http://www.hit.uib.no/corpora/]). In the context of OLAC, the term *language resource* is broadly construed to include all three of these: data, tools and advice.

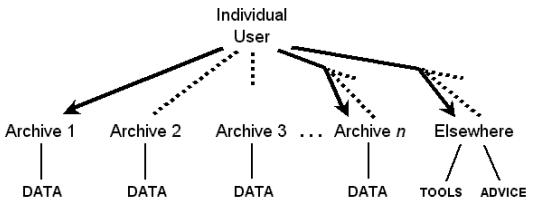

Figure 1: In reality the user can't always get there from here

Unfortunately, today's user does not have ready access to the resources that are needed. Figure 1 offers a diagrammatic view of this reality. Some archives (e.g. Archive 1) do have a site on the internet which the user is able to find, so the resources of that archive are accessible. Other archives (e.g. Archive 2) are on the internet, so the user could access them in theory, but the user has no idea they exist so they are not accessible in practice. Still other archives (e.g. Archive 3) are not even on the internet. And there are potentially hundreds of archives (e.g. Archive n) that the user needs to know about. Tools and advice are out there as well, but are at many different sites.

There are many other problems inherent in the current situation. For instance, the user may not be able to find all the existing data about a language of interest because different sites have called it by different names (low *recall*). The user may be swamped with irrelevant resources because search terms have important meanings in other domains (low *precision*). 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 have no basis for judging its merits.

As web-indexing technologies improve one might hope that a general-purpose search engine should be sufficient to bridge the gap between people and the resources they need. However this is a vain hope. First, many language resources, such as audio files and software, are not text-based. Second, many language names have several variants, and these various strings regularly denote things other than languages. Third, much of the material is not–and will never be–documented in free prose on the web. In place of traditional web-indexing, two new initiatives provide the necessary infrastructure for language resource discovery.

The Dublin Core Metadata Initiative began in 1995 to develop conventions for resource discovery on the web [dublincore.org]. The Dublin Core metadata elements represent a broad, interdisciplinary consensus about the core set of elements that are likely to be widely useful to support resource discovery. The Dublin Core consists of 15 metadata elements, where each element is optional and repeatable: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage and Rights. This set can be used to describe resources that exist in digital or traditional formats.

The Open Archives Initiative (OAI) was launched in October 1999 to provide a common framework across electronic preprint archives, and it has since been broadened to include digital repositories of scholarly materials regardless of their type.[1] To implement the OAI infrastructure, an archive must comply with two standards: the OAI *Shared Metadata Set* (Dublin Core), which facilitates interoperability across all repositories participating in the OAI, and the OAI *Metadata Harvesting Protocol*, which allows software services to query a repository using HTTP requests.

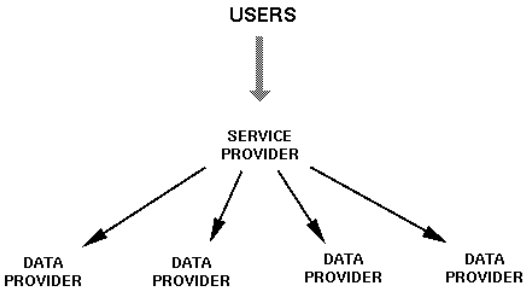

Figure 2: A Service Provider Accessing Multiple Data Providers

OAI archives are called "data providers," and typically have a submission procedure, a long-term storage system, and a mechanism permitting users to obtain materials from the archive. An OAI "service provider" provides end-user services–such as search functions over union catalogs–based on metadata harvested from one or more data providers. Figure 2 illustrates a single service provider accessing three data providers using the OAI metadata harvesting protocol. End-users only interact with service providers.

The OAI infrastructure has the bottom-up, distributed character of the web, while simultaneously having the efficient, structured nature of a centralized database. This combination is well-suited to the language resource community, where the available data is growing rapidly and where a large user-base is fairly consistent in how it describes its resource needs.

OAI data providers may support metadata standards in addition to the Dublin Core. Thus, a specialist community like the language resources community can define a metadata format tailored to its domain. Using the OAI infrastructure, the community's archives can be federated: a virtual meta-archive collects all the information into a single place and end-users can query multiple archives simultaneously. In the case of OLAC, the Linguistic Data Consortium has harvested the catalogs of ten participating archives and created a search interface which permits queries over all 9,000+ records. A single search typically returns records from multiple archives. The prototype can be accessed via [www.language-archives.org].

[1] www.openarchives.org; (Lagoze and de Sompel, 2001)

2 A Core Metadata Set for Language Resources

The OLAC Metadata Set extends the Dublin Core set only to the minimum degree required to express those basic properties of language resources which are useful as finding aids. All fifteen Dublin Core elements are used in the OLAC Metadata Set. In order to suit the specific needs of the language resources community, the elements have been "qualified" following principles articulated in "Dublin Core Qualifiers" (DCMI, 2000), and explained below. This section lists the OLAC metadata elements and controlled vocabularies. Full details are available from [www.language-archives.org].

2.1 Metadata attributes

Three attributes – refine, code, and lang – are used throughout the metadata set to handle most qualifications to Dublin Core. Some elements in the OLAC Metadata Set use the refine attribute to identify element refinements. These qualifiers make the meaning of an element narrower or more specific. A refined element shares the meaning of the unqualified element, but with a more restricted scope (DCMI, 2000).

Some elements in the OLAC Metadata Set use the code attribute to hold metadata values that are taken from a specific encoding scheme. When an element may take this attribute, the attribute value specifies a precise value for the element taken from a controlled vocabulary or formal notation (§2.3). In such cases, the element content may also be used to specify a free-form elaboration of the coded value.

Every element in the OLAC Metadata Set may use the lang attribute. It specifies the language in which the text in the content of the element is written. The value for the attribute comes from a controlled vocabulary OLAC-Language. By default, the lang attribute has the value "en", for English. Whenever the language of the element content is other than English, the lang attribute should be used to identify the language. By using multiple instances of the metadata elements tagged for different languages, data providers may offer their metadata records in multiple languages. Service providers may use this information in order to offer multilingual views of the metadata.

2.2 The elements of the OLAC Metadata Set

In this section we present a synopsis of the elements of the OLAC metadata set. Some elements are associated with a controlled vocabulary. These are parenthesized and discussed later. Each element is optional and repeatable.

Title: A name given to the resource.

Creator: An entity primarily responsible for making the content of the resource (OLAC-Role).

Subject: The topic of the content of the resource.

Subject.language: A language which the content of the resource describes or discusses (OLAC-Language).

Description: An account of the content of the resource.

Publisher: An entity responsible for making the resource available.

Contributor: An entity responsible for making contributions to the content of the resource (OLAC-Role).

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

Type: The nature or genre of the content of the resource (DC-Type).

Type.linguistic: The nature or genre of the content of the resource, from a linguistic standpoint (OLAC-Linguistic-Type).

Type.functionality: The functionality of a software resource (OLAC-Functionality).

Format: The physical or digital manifestation of the resource (OLAC-Format).

Format.cpu: The CPU required to use a software resource (OLAC-CPU).

Format.encoding: An encoded character set used by a digital resource (OLAC-Encoding).

Format.markup: The OAI identifier for the definition of the markup format (OLAC-Markup).

Format.os: The operating system required to use a software resource (OLAC-OS).

Format.sourcecode: The programming language(s) of software distributed in source form (OLAC-Sourcecode).

Identifier: An unambiguous reference to the resource within a given context (e.g. URI, ISBN).

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

Language: A language of the intellectual content of the resource (OLAC-Language).

Relation: A reference to a related resource.

Coverage**:** The extent or scope of the content of the resource (e.g. spatial or temporal).

Rights**:** Information about rights held in and over the resource (OLAC-Rights).

Rights.software**:** Information about rights held in and over a software resource. (OLAC-Software-Rights).

Observe that some elements, such as Format, Format.encoding and Format.markup are applicable to software as well as to data. Service providers can exploit this feature to match data with appropriate software tools.

2.3 The controlled vocabularies

Controlled vocabularies are enumerations of legal values for the code and refine attributes, and are currently undergoing development. In some cases, more than one value applies and the corresponding element must be repeated, once for each applicable value. In other cases, no value is applicable and the corresponding element is simply omitted. In yet other cases, the controlled vocabulary may fail to provide a suitable item, in which case the most similar vocabulary item can be optionally specified, and a prose comment included in the element content.

OLAC-Language: A vocabulary for identifying the language(s) that the data is in, or that a piece of linguistic description is about, or that a particular tool can process.

OLAC-Linguistic-Type: The primary linguistic descriptors for a language resource: transcription, annotation, description and lexicon (with subcodes for each type).

OLAC-CPU: A vocabulary for identifying the CPU(s) for which the software is available, in the case of binary distributions: x86, mips, alpha, ppc, sparc, 680x0.

OLAC-Encoding: A vocabulary for identifying the character encoding used by a digital resource, e.g. iso-8859-1, ...

OLAC-Format: A vocabulary for identifying the manifestation of the resource. The representation is inspired by MIME types, e.g. text/sf for SIL standard format. (Format.markup is used to identify the particular tagset.) It may be necessary to add new types and subtypes to cover non-digital holdings, such as manuscripts, microforms, and so forth and we expect to be able to incorporate an existing vocabulary.

OLAC-Functionality: A vocabulary for classifying the functionality of software, again using the MIME style of representation, and using the HLT Survey as a source of categories (Cole, 1997) as advocated by the ACL/DFKI Natural Language Software Registry. For example, written/OCR would cover "written language input, print or handwriting optical character recognition."

OLAC-OS: A vocabulary for identifying the operating system(s) for which the software is available: Unix, MacOS, OS2, MSDOS, MSWindows. Each of these has optional subtypes, e.g. Unix/Linux, MSWindows/winNT.

OLAC-Rights: A vocabulary for classifying the rights held over a resource, e.g.: open, restricted, ...

OLAC-Role: A vocabulary for identifying the role of a contributor or creator of the resource, e.g.: author, editor, translator, transcriber, sponsor, ...

OLAC-Software-Rights: A vocabulary for classifying the rights held over a resource, e.g.: open-source, royalty-free-library, royalty-free-binary, commercial, ...

OLAC-Sourcecode: A vocabulary for identifying the programming language(s) used by software which is distributed in source form, e.g.: C++, Java, Python, Tcl, VB, ...

3 Issues for the Asian Language Resources Community

Language identification is probably the most fundamental kind of information that can be given to any language resource (Simons, 2000). The most comprehensive knowledge base for language identification is the Ethnologue (Grimes, 2000), an online searchable database which has been built up over fifty years. The Ethnologue database contains several types of information for each language: a unique three-letter code, the country where this language is spoken, alternative names, dialects, language classification, comments, and references to the SIL bibliography. This section discusses a variety of issues relating to language identification and to the specification of the languages covered by multilingual resources. We believe these are two key issues for the Asian language resources community.

3.1 Issues with language identification

In order to identify and catalog a language it is crucial to define what counts as a language and to distinguish languages from dialects. The editors of the Ethnologue have made thousands of such decisions using advice from hundreds of experts around the world. However, for many languages scholarship remains patchy or else there is scholarly disagreement. In such cases, the best that the Ethnologue can do is what it does already– represent incomplete knowledge and then produce periodic updates to reflect the results of new research.

As OLAC grows, Ethnologue codes will be deployed widely. Each new OLAC-conformant archive will be faced with a range of issues in seeking to associate language codes with language resources. For concreteness, we have chosen for our examples the Formosan languages, a group of Austronesian languages spoken in Taiwan. We put ourselves in the shoes of the field research group at Academia Sinica (Elizabeth Zeitoun, personal communication) and try to envisage the problems which they might encounter in assigning Ethnologue codes to their language resources.

We see three broad categories of problem: over-splitting, over-chunking and omission. Over-splitting occurs when a language variety is treated as a distinct language. For example, Nataoran is given its own language code (AIS) even though the scholars at Academia Sinica consider it to be a dialect of Amis (ALV). Over-chunking occurs when two distinct languages are treated as dialects of a single language (there does not appear to be an example of this in the Ethnologue's treatment of Formosan languages). Omission occurs when a language is not listed. For example, two extinct languages, Luilang and Quaquat, are not listed in the Ethnologue. Another kind of omission problem occurs when the language is actually listed, but the name by which the archivist knows it is not listed, whether as a primary name or an alternate name. In such a case the archivist cannot make the match to assign the proper code. For instance, the language listed as Taroko (TRV) in the Ethnologue is known as Seediq by the Academia Sinica; several of the alternate names listed by the Ethnologue are similar, but none matches exactly.

Beyond these three problems with language identification, a further type of problem concerns scholarly disagreement over language family classification. The Ethnologue follows the Oxford International Encyclopedia of Linguistics (Bright, 1992) for most language families. For the Austronesian languages, including the Formosan languages, the Ethnologue follows the Comparative Austronesian Dictionary (Tryon, 1994). Additionally, some changes have been entered in the light of more recent comparative studies.[2] Academia Sinica has developed its own language family classification scheme for Formosan languages, and this differs from the Ethnologue/Tryon scheme. Additionally, languages typically have many variant names, and scholars may disagree on the choice of a canonical name for the language. For example, the Academia Sinica scholars believe Taroko to be a variant of Seediq, while the Ethnologue/Tryon would presumably consider Seediq to be a variant of Taroko.

The consequences of these problems for classification and retrieval are obvious. In the case of over-splitting, as with AIS and ALV mentioned above, someone searching for Amis resources will need to know to search over both codes. An archivist cataloging a resource which is ambiguous with respect to the AIS/ALV distinction (perhaps because it was created by someone who did not believe in the distinction) may need to assign both codes. In the case of over-chunking, an archivist cannot specify the individual language but must use a code which designates two or more languages. Someone searching for resources in one of those languages will experience lower precision. In the case of omission, no language code can be assigned, and classification and search must fall back to using conventional string representations for language names (with the attendant precision and recall problems). In the case of differing language family classifications, the precision and recall of searches on language family names are reduced.

All of these problems can be addressed through existing Ethnologue mechanisms.[3] However, OLAC metadata and service providers could offer complementary remedies.

Controlling element content. The Language and Subject.language elements permit the

[2] More information is available online at http://www.ethnologue.com/ethno_docs/introduction.asp. The Formosan language family can be viewed at http://www.ethnologue.com/show_family.asp?subid=982.

[3] See the four questionnaires at http://www.ethnologue.com/ethno_docs/questionnaires.asp.

language code to be specified in the code attribute, while the element content is unrestricted. A community of Formosan scholars could develop a controlled vocabulary for identifying speech varieties down to any level of detail they liked, and then use those terms as the content of the Language or Subject.language element. For example, the following are five varieties of the Bunun language:

```
<language code="x-sil-BNN">Northern/Takituduh</>
<language code="x-sil-BNN">Northern/Takibakha</>
<language code="x-sil-BNN">Central/Takbanuaz</>
<language code="x-sil-BNN">Central/Takivatan</>
<language code="x-sil-BNN">Southern/Isbukun</>
```

If no Ethnologue code corresponded to the group of languages in question, as in the Amis/Nataoran case, the code attribute could be omitted (though this would prevent recall on the Ethnologue code). This general approach could be formalized by permitting subcommunities to register an encoding scheme as a controlled vocabulary with a unique name. That name would be specified as the value of a new scheme attribute, and the element content would be constrained to be an item from the corresponding vocabulary. These approaches would address the problems of over-chunking and omission.

Registering language groups with an OLAC registration service. While the classification of a language is sometimes treated as metadata for resources in that language, we believe that a more appropriate location for this type of finding aid is in OLAC service providers. OLAC could maintain a language classification server which would house a comprehensive list of language family names and their extensional definitions (i.e. sets of Ethnologue codes). The server would permit users to define their own language group names or their own versions of existing group names. For instance, Academia Sinica could register a language group name AS:Amis with the extension {ALV, AIS}. Searching on their notion of "Amis" would return resources classified under both codes. Entire classification schemes with complex hierarchies could be represented in this fashion. OLAC service providers could index their harvested metadata using these names, allowing any user to perform searches using any classification scheme. Over time, the more respected and popular classifications could be identified and accorded due prominence. This mechanism would address the problems of over-splitting and differing classification.

3.2 Issues with multilingual resources

For many language resources it is necessary to identify more than one language. In some cases, such as MT systems and bilingual lexicons, there is an added complication, namely *directionality*. Someone looking for such a resource will usually want to specify source and/or target languages. For example, in searching for a Korean-to-English resource, the user would not usually be interested in discovering English-to-Korean resources. Thus, there is an *a priori* need for resources to be described using *ordered pairs* of languages.

OLAC provides some mechanisms which can be applied to these cases. First, OLAC metadata elements are repeatable, e.g. a system which can process multiple languages will be described using multiple Subject.language elements, one per language. Second, OLAC incorporates directionality in its distinction between Language and Subject.language. Are these simple mechanisms adequate for the convenient and accurate description and discovery of multilingual resources? In this section we enumerate the main types of multilingual resource and show how OLAC metadata can be used to classify them. We report some problems and discuss possible solutions.

Machine Translation Systems

The simplest type of MT system is a unidirectional system which translates language S to language T. Here, the intended audience of such a system is assumed to be the speakers of language T who use the system to access documents in another language S. The OLAC solution would be to designate S as the Subject.language and T as the Language.

Note that "audience" is slightly problematic. Such an MT system may be intended for an audience of S speakers who wish to translate their documents into language T. The problem here is not with directionality but with the notion of "audience" in the OLAC definition of Language. The definition could be adjusted to remove this problem.

Next in order of complexity is the bidirectional case, where a system translates in both directions between languages X and Y. Extending the previous solution, we would designate both X and Y as Language and Subject.language. Ideally, we would use order or structure to group the languages appropriately:

```
<pair><Subject.language code=X/>
      <Language code=Y/></pair>
<pair><Subject.language code=Y/>
      <Language code=X/></pair>
```

However, OLAC metadata is flat and unordered. The only available options are permutations of the following, in which we can make no contrastive use of order.

```
<Language code=X/>
<Language code=Y/>
<Subject.language code=X/>
<Subject.language code=Y/>
```

Although this loses information, we do not believe it presents a problem for typical kinds of retrieval. Queries for an MT system (i) from X; (ii) from Y; (iii) to X; (iv) to Y; (v) from X to Y; or (vi) from Y to X, will discover the system described above.

Next are MT systems which translate from one language into many, or from many languages into one (star configurations). Here the obvious approach is adequate:

One-to-many:

```
<Subject.language code=S/>
<Language code=T1/>
<Language code=T2/>
<Language code=T3/>
```

Many-to-one:

```
<Subject.language code=S1/>
<Subject.language code=S2/>
<Subject.language code=S3/>
<Language code=T/>
```

Finally, there are MT systems which translate *from* and *to* all languages in a set of n languages. Here again the obvious approach is adequate, and is clearly superior to a solution where all $n(n-1)$ ordered pairs are enumerated.

```
<Subject.language code=X/>
<Subject.language code=Y/>
<Subject.language code=Z/>
<Language code=X/>
<Language code=Y/>
<Language code=Z/>
```

Multilingual Lexicons

A unidirectional bilingual lexicon with lemmas in language S and definitions in language T is described like a unidirectional MT system. The OLAC solution is to designate S as the Subject.language and T as the Language. A fully bidirectional lexicon intended for use by speakers of either language would be described in the same fashion as a bidirectional MT system:

```
<Language code=X/>
<Language code=Y/>
<Subject.language code=X/>
<Subject.language code=Y/>
```

As before, this solution has lost some structure, but retrieval behavior is correct nonetheless. Of course, this metadata could equally be the collection-level metadata for a set of two monolingual dictionaries, one in language X and one in language Y. However, the Type.linguistic element would distinguish these cases by having different vocabulary items for monolingual and bilingual lexicons.

Multilingual lexicons may also exhibit star configurations: one-to-many (lexicons with definitions in multiple languages); or many-to-one (comparative wordlists). Here the treatment is analogous to the corresponding MT systems discussed above.

Finally, multilingual lexicons may map between all pairs of a set of languages, as in the case of some termbanks. In this case, all the languages are designated both using Language and Subject.language.

Text Collections

The most simple case of a text collection is a set of texts in a single language. Usually, the language in these texts would be described using the Language element. However, in the situation where the text collection is intended to document a language, then it is simultaneously *in* and *about* that language. Accordingly, it would be tempting to describe this with both the Language and Subject.language elements. Rather than let the decision about metadata depend on the intent of the creator of the resource (which may not be known), or on the typical usage of the resource (which may change through time), we think it would be simplest to describe such situations using the Language element only. This approach generalizes to the case of a text collection spanning multiple languages (e.g. where each text comes from one of the Formosan languages). Here, collection-level metadata would provide a Language element for each language represented in the collection.

In the case of bilingual aligned texts (bitexts) there is normally a directionality, since one of the texts is primary and the other is a translation. Here, we specify the primary language using Subject.language, and the translation using Language.

A complication arises when the texts or bitexts employ analytical notations. For a text in language X transcribed in a notation (such as the International Phonetic Alphabet) which is inaccessible to speakers of X (assuming any exist), it

would make sense to use Subject.language instead of Language. The situation becomes more vexing for texts with embedded annotations (such as the notation of Conversation Analysis), where non-specialists could discard the annotations to get a conventional text. Here, the use of both Subject.language and Language seems to be indicated.

Descriptions

The final category we consider is linguistic descriptions, such as field notes and grammars. In the usual case, the language described is specified with Subject.language while the language of the commentary is specified with Language. One interesting case is where a third language is used for elicitation. For example, a sentence from Amis may have been elicited using a sentence from Chinese, and both sentences may have been entered in the field notes. Next, commentary in the language of the linguist, such as English, may have been added. In this case, we would say that the field notes include bitext, and the languages of the bitext would be described in the usual way. The audience language of the field notes would also be specified. Using OLAC's flat metadata, we would specify the languages as follows:

```
<Subject.language code="x-sil-AIS"/>
<Language code="x-sil-CFR"/>
<Language code="x-sil-ENG"/>
```

We believe this is perfectly adequate for the majority of retrieval purposes. If it were necessary to represent the structure more accurately, *two* OLAC records could be associated with the same resource, one describing the field notes as a whole (with the above language designations), and one describing the bitext content (with just the AIS and CFR designations). The different linguistic types would be expressed using the Type.linguistic element, and the two records could refer to each other using the Relation element:

```
<Relation refine="isPartOf">id1</Relation>
<Relation refine="hasPart">id2</Relation>
```

4 Conclusions

This paper has presented the leading ideas of the Open Language Archives Community (OLAC), along with its metadata set and controlled vocabularies. Language resources exhibit great diversity, and include all types of data, tools and advice. As collections of language resources proliferate, OLAC will make it maximally easy for members of the language resources community to discover each other's resources. Another dimension of OLAC, not discussed here, will permit community-agreed best practices to be identified, greatly facilitating resource re-use.

Members of the Asian language resources community are encouraged to join OLAC and contribute to the development of OLAC metadata, vocabularies, and archives.

Acknowledgments

The work reported here is supported by the National Science Foundation under grants: 9910603 *International Standards in Language Engineering*, 9978056 *TalkBank*, 9983258 *Linguistic Exploration*, and by a Taiwanese National Digital Archives Pilot project *Chinese and Austronesian Corpora*. The LDC prototype was developed by Éva Bánik.

References

[Bird and Simons2001] Steven Bird and Gary Simons. 2001. The OLAC metadata set and controlled vocabularies. In *Proceedings of ACL/EACL Workshop on Sharing Tools and Resources for Research and Education*. http://arXiv.org/abs/cs/0105030.

[Bright1992] William Bright, editor. 1992. *International Encyclopedia of Linguistics*. Oxford University Press.

[Cole1997] Ronald Cole, editor. 1997. *Survey of the State of the Art in Human Language Technology*. Studies in Natural Language Processing. Cambridge University Press. http://cslu.cse.ogi.edu/HLTsurvey/.

[DCMI1999] DCMI. 1999. Dublin Core Metadata Element Set, version 1.1: Reference description. http://dublincore.org/documents/1999/07/02/dces/.

[DCMI2000] DCMI. 2000. Dublin Core qualifiers. http://dublincore.org/documents/2000/07/11/dcmes-qualifiers/.

[Grimes2000] Barbara F. Grimes, editor. 2000. *Ethnologue: Languages of the World*. Dallas: Summer Institute of Linguistics, 14th edition. http//www.sil.org/ethnologue/.

[Lagoze and de Sompel2001] Carl Lagoze and Herbert Van de Sompel. 2001. The Open Archives Initiative: Building a low-barrier interoperability framework. http://www.cs.cornell.edu/lagoze/papers/oai-jcdl.pdf.

[Simons2000] Gary Simons. 2000. Language identification in metadata descriptions of language archive holdings. In Steven Bird and Gary Simons, editors, *Proceedings of the Workshop on Web-Based Language Documentation and Description*. http://www.ldc.upenn.edu/exploration/expl2000/papers/simons/.

[Tryon1994] Darrell T. Tryon, editor. 1994. *Comparative Austronesian Dictionary: An Introduction to Austronesian Studies*. Number 10 in Trends in Linguistics: Documentation. Walter De Gruyter.