

Lev Michael
Department of Anthropology
University of Texas at Austin
Austin, TX 78712-1086
lmichael@mail.utexas.edu

Abstract

This paper introduces the proposal by the Archive for the Indigenous Languages of Latin America (AILLA) group to include the following ten metadata elements in the metadata standard for the linguistic database community: Recorder, Subject Language, Commentary Language, Notable Participants, Register/Style, Channel, Genre, Event, Transcription, Translation.

These semantics of these metadata elements are discussed, as are some of the broader principles that guided their creation. A number of distinctions used by AILLA in the development of these metadata elements, including global versus featural metadata and data-external versus data-internal metadata, are also examined.

A brief comparison of the AILLA metadata scheme with other metadata standards, among them MARC, TEI, and CGAR, is included.

One possible solution to the requirement for comprehensive, modifiable, and interoperable metadata schemes is discussed: the formation of an institutional body for the maintenance of metadata structures and standards.

Cells of Infrastructure Table Addressed: Metadata, Store

Title: Creating Discourse Data Metadata for the AILLA Project: Lessons Learned and Needs Felt

INTRODUCTION

This paper introduces the proposal made by the Archive of the Indigenous Languages of Latin America (AILLA) group, for ten metadata elements to be considered for inclusion in the metadata standard for the linguistic database community. I will begin by describing aspects of the metadata scheme devised by the AILLA group, since the current proposal is based on certain guiding principles embodied in our metadata scheme, and because we believe that some of these principles can be profitably discussed in reference to the metadata standard. I will then describe the elements in our metadata proposal, and briefly compare these elements, and the principles behind them, to those found in other metadata schemes, including MARC, TEI, and CGAR. I conclude with an enumeration of issues that the AILLA group feels should be addressed in our discussions here.

The AILLA group, led by Joel Sherzer, Tony Woodbury, and Mark McFarland of the University of Texas, seeks to create a permanent web-accessible database of digitized audio and textual materials drawn from the indigenous languages of Latin America, focusing especially on naturally occurring discourse.

The primary purpose of the AILLA metadata scheme is to provide information about materials held by AILLA that will enable users to find resources relevant to their research interests. The majority of AILLA metadata is designed to be used with search interfaces that will allow users to search over metadata elements for element types, or combinations of element types joined by Boolean operators. The elements and element types are oriented towards language- and discourse-oriented research, and many are based on taxonomies and classification systems employed by linguists and linguistic anthropologists.

AILLA metadata differs in this way from, for example, MARC metadata, which has no principal disciplinary focus. We believe that creating a metadata standard for the linguistic database community that explicitly reflects and is shaped by the research interests of the community has much to recommend it, and we hope that in our discussions the advantages of disciplinary focus will be given consideration.

Due to the nature of the materials we are archiving, we anticipate that linguists and linguistic anthropologists will form the bulk of users. However, a touchstone in our thinking about metadata was making materials in the AILLA database usefully searchable for scholars in a wide number of disciplines. Properly organized, we believe that AILLA will be of interest not just to linguists and linguistic anthropologists, but to cultural anthropologists, ethnomusicologists, folklorists, historians, and area-studies specialists in other disciplines as well. We hope that in elaborating the metadata standard, the issue of cross-disciplinary utility will be kept in mind, since although we may gather or archive materials as ‘linguistic’ materials, the lives of these materials in the scholarly (and non-scholarly) world may be wider than we might imagine.

Clearly the two issues just raised, the value of disciplinary focus in metadata and the value of cross-disciplinary utility in metadata, are in tension with one another.

Another relevant aspect of the AILLA metadata scheme is the approach we have taken to structuring the metadata. Rather than choosing a small number of very broad metadata elements to characterize the resources in the AILLA database, each with a large number of possible types, we have instead opted to construct numerous orthogonal axes of classification (or at least nominally orthogonal ones - I will return to this below) in which each axis encompasses a fairly modest number of distinctions (modest = <40).

By orthogonal, I mean that the distinctions made along a given axis of classification (that is, the element types of a given element) are independent from those made along another. To jump ahead for a moment to some of the metadata elements we are proposing for the metadata standard, consider channel, genre, and speech event. These metadata elements are orthogonal, since the element type distinctions they encompass can freely combine:

Channel	Genre	Speech Event	Example
chanting	political oratory	community meeting	Kuna chiefs’ political oratory
speaking	political oratory	funeral ceremony	Pericles’ funeral oration

speaking	prayer	funeral ceremony	Protestant funeral prayer
singing	prayer	funeral ceremony	Jewish funeral prayer

We have several reasons for structuring our metadata in this way. First, we wanted to enable the user to easily obtain a sense of the search space. By being able to see (via pull-down menus, for example) the ‘axes’ of the search space, and the distinctions made along those axes, the user is quickly able to understand the classification system employed by AILLA, and tune his or her search activities to this system. This, we hope, will enable users to avoid the problem associated with searching via keywords of being unsure how to look for a topic of interest due to the sheer vastness of the set of (unstructured) keywords.

It is also our aim that the kinds of searches we anticipate scholars to make in the AILLA database will be easy to execute. We can easily tailor our metadata scheme to the needs of archive users and enable them to carry out searches that combine or exclude element types. Using standard keyword searches that use common thesauri (e.g. Library of Congress), for example, it would difficult to search a database for “chanted dialogical greetings,” an important discursive practice in the northern Amazon and adjacent areas. We chose the distinctions on which we based our metadata by reviewing the scholarly literature on indigenous Latin American discourse, and by supplementing this information with knowledge that the members of the AILLA group, and advisors we consulted, have of this topic.

Another reason for organizing our metadata in this way was to provide depositors a convenient framework for describing the materials they archive. By providing this framework we hope to encourage depositors to provide a useful minimum of information about deposited materials, and at the same time to avoid idiosyncratic variation that would make such information of little use to other researchers. It is worth noting that by providing this metadata, depositors are in effect contributing to an areal-typological and comparative analysis of indigenous Latin American discourse.

In thinking about what to include in, and exclude from, our metadata, we developed several distinctions that we found “useful for thinking with” that might also be useful for us here.

One such distinction was between data-internal and data-external metadata. The terms “data-internal” and “data-external” serve to distinguish metadata that refers to features found in the data itself, such as the subject language and channel, from features that are not found in the data itself, but could rather be called “contextual information”, such as place, time, and speech event.

Another distinction we found useful was global versus featural metadata. The distinction here is between metadata that characterizes a stretch of discourse as a whole versus metadata that characterizes small portions or features of the discourse. Metadata that specifies what genre a piece of discourse belongs to is a prototypical global metadata category, while metadata that tags instances of conversational overlap would be a good example of featural metadata.

Since, for AILLA, it is the depositor who supplies the vast majority of the metadata, we have tended towards data-internal, global metadata, as most researchers tend not to systematically gather data on communicative context, and because most language materials are not (yet) tagged for featural information. We felt that requiring substantial data-external and featural metadata would place to great a burden on depositors.

THE AILLA METADATA PROPOSAL

Our contribution to the metadata standard is a proposal for a set of ten metadata elements (see attachment): **Recorder, Subject Language, Commentary Language, Notable Participants, Register/Style, Channel, Genre, Event, Transcription, Translation.**

The first three elements (**Recorder, Subject Language, Commentary Language**) are intended as refinements of Dublin Core metadata elements. The first, **Recorder**, which identifies the individual who recorded (by audio, video, or writing) the resource, is intended to draw a distinction between the **Creator** of a resource, who is the entity primarily responsible for the content of the resource, and the individual who is responsible for that content being recorded on some durable medium. This distinction is important in fieldwork contexts.

The next two elements, **Subject Language** and **Commentary Language**, are intended as refinements of the Dublin Core **Language** element. The **Subject Language** is the data language, or the language which is the object of analysis, whereas the **Commentary Language** is the language in which the analysis is carried out. By using these two elements, for example, it will be possible to easily distinguish between, for example, a grammar of Russian, written in English, and a grammar of English, written in Russian.

The next five elements, **Notable Participants, Register/Style, Channel, Genre, and Event**, are categories relevant to socially- and performance-oriented research on language. Moreover, these elements encode information that will be useful to researchers in other disciplines. The permitted element types for each element are drawn from a modest list.

The last two elements, **Transcription** and **Translation**, pertain to ways in which linguistic data is processed (and to some degree, analyzed). Since transcription strategies and forms of translation can vary quite considerably, and their utility for particular research projects is similarly variable, we believe that it is useful for users to be able to specify the types of transcription and translation they are interested in. The permitted transcription and translation types are drawn from a modest list.

It is worth pointing out that we do not mean these metadata elements to make iron-clad ontological distinctions; rather, these metadata are intended as resources for search tools. We found this distinction to be an important one. Drawing a solid ontological distinction between some elements, and among some element types, can be difficult. For example, attempting to draw a clear distinction between folk tales and myths involves many thorny theoretical and empirical issues. Perhaps more difficult is attempting to decide definitively how particular pieces of discourse fit into what are, ultimately, etic categories. These problems largely dissolve, though, when one realizes what is at stake is not whether any stretch of discourse is

fundamentally a myth or a folk tale, and what these categories precisely distinguish, but rather, if the categories will serve as useful reference points for searches carried out by users. Thus, troubling questions like what precisely constitutes a register or style can be left partially answered.

COMPARISONS WITH OTHER METADATA SCHEMES

I want to draw some brief contrasts between AILLA metadata and MARC (Machine Readable Cataloging) metadata, TEI (Text Encoding Initiative) metadata, and CGAR (Council for the Preservation of the Anthropological Record (CoPAR) Guide to Anthropological Records) metadata, which are among the widely used metadata schemes that might be relevant to the task of characterizing linguistic materials.

First, it should be noted that none of these metadata schemes is designed for linguistically-oriented research. The CGAR (CoPAR Guide to Anthropological Records) metadata scheme is oriented towards anthropological resources, but since it ultimately relies on extant thesauri (Library of Congress, etc.) for keyword lists that are used to specify very general elements (e.g. **Domain**, which can refer to the relevant, culture, language, time period, etc.) it is not particularly more suited to searches for anthropologically- or linguistically-relevant materials than MARC metadata.

Second, none of these metadata schemes allow for the kind of multi-dimensional characterization of resources that the AILLA metadata scheme aims at, making searches for specific materials of interest a difficult affair, since it is not possible to characterize with precision the type of resource one is interested in.

TEI, as the name suggests, is substantially aimed at tagging text - what we have called data-internal metadata. Thus, it has not been an important resource in constructing our own metadata, although we imagine that it might serve as a useful starting point for discussion about aspects of the metadata standard that have not been of direct concern to the AILLA group.

ISSUES FOR DISCUSSION

There are contingent issues raised by the metadata proposal we are making here. For example, if the metadata elements we are proposing are indeed incorporated into the metadata standard, it would be useful to discuss broadening the element types we enumerate (e.g. genre types), which are especially focused on discourse types in indigenous Latin American cultures. We imagine, for example, that the LACITO project, which focuses on other parts of the world, would find our element types too restrictive, and would require additions relevant for the cultures and languages of their areas.

Similarly, the metadata elements we have proposed clearly reflect our research interest in comparative studies of discourse processes, and our focus on discourse as a social phenomenon. The materials we are archiving could no doubt be profitably described with metadata relating to phonetic, phonological, morphological, syntactic, or semantic features. AILLA would find obtaining such metadata difficult, but it is not hard to imagine that a group of phoneticians would

find such metadata both crucial and easy to provide for data they gathered. This raises the issue of what degree of (sub)disciplinary specificity we wish to entertain in our metadata, a topic for further discussion. Should the metadata standard be sufficiently broad and detailed that it is capable of encoding distinctions and features relevant to all possible disciplines and subdisciplines that are likely to take advantage of the databases which employ this standard? Or should instead the metadata standard seek only to encode features of such generality that no database manager will have difficulty providing metadata that conforms to the standard, such as for example, Dublin Core does? More sensibly, where does the balance lie between these two extremes?

Finally, it is reasonable to expect that it will be necessary for the metadata standard to evolve over time. I suspect this will be especially true in the early formation period, as the metadata standard aligns with the practices of the linguistic community. Over the long-term too, the metadata standard will need to evolve as linguistic practice itself changes. Many of the distinctions that we draw in the AILLA metadata scheme, for example, arise from research in the ethnography of communication tradition, and we may not have made these same distinctions had we been carrying out this same project in the 1950s.

This suggests that even after a metadata standard is developed, the need for an institutional body to oversee the long-term support of a metadata standard for the linguistic database community will remain. It can be hoped that one of the fruits of our work here will be such an institutional framework.

APPENDIX: PROPOSED METADATA ELEMENTS

1. Element: Recorder

Name: Recorder

Identifier: Recorder

Definition: Name or identification of person who recorded, in either text, audio, or video formats, the content of the resource.

Comment: This allows a distinction to be drawn between the creator of the resource (the "entity primarily responsible for the content of the resource"), and the person who recorded, preserving the content for archiving or study. Thus a resource that is a recording of a political speech would have as its Creator the politician, but as its Recorder the researcher who made the recording.

2. Element: Subject Language

Name: Subject Language

Identifier: Subject Language

Definition: Name of the language in the resource that is the subject of annotation, commentary, or analysis.

Comment: Instead of two-letter ISO codes, we recommend as best practice the use of the most recent three-letter Ethnologue codes. The distinction between subject language and commentary language (see next element) is not made by Dublin Core. Using the element "Language" in Dublin core, there is no natural way to distinguish between, for example, an analysis of English grammar, written in Russian, and an analysis of Russian grammar, written in English.

3. Element: Commentary Language

Name: Commentary Language

Identifier: Commentary Language

Definition: Name of the language that is used to comment on, annotate or analyze the subject language of the resource.

Comment: See comments for "Subject Language"

4. Element: Notable Participants

Name: Notable Participants

Identifier: Notable Participants

Definition: Culturally or socially notable individual whose verbal or physical behavior forms part of the content of the resource.

Comment: See Notable Participant Types below.

Notable Participant Types

Political leader/figure

Medicinal specialist

Ritual specialist
School teacher
Storyteller
Researcher
Consultant
Translator
Interpreter
Patient

5. Element: Register/Style

Name: Register/Style

Identifier: Register/Style

Definition: Any specialized or restricted language employed by speakers.

Comment: See List of Register/Style Types below.

Register/Style Types

Code Switching
Honorific Speech
Specialist's Language
Esoteric Speech
Play Language
Baby/Caretaker Talk
Joking
Formal Speech
Informal/Conversational Speech
Nonsense/Unintelligible Speech

6. Element: Channel

Name: Channel

Identifier: Channel

Definition: Acoustically distinctive communicative modality.

Comment: See list of Channel Types below

Channel Types

Whispering
Muttering
Talking
Singing

Falsetto
Chanting
Wailing
Shouting
Musical Instrument

7. Element: Genre

Name: Genre

Identifier: Genre

Definition: Culturally salient and/or structurally distinguishable forms of communicative interaction.

Comment: See list of Genre Types below.

Genre Types

Conversation
Greeting
Leave Taking
Interview
Linguistic Elicitation
Personal Narrative/Story
Traditional Narrative/Story
Historical Narrative/Story
Folktale
Humorous Story
Trickster Story
Myth
Proverb
Riddle
Joke
Dream Report
Divinatory Speech
Report
Advice/Counsel
Political Oratory
Religious Oratory
Argument
Announcement
Hortatory Speech
Insult
Verbal Dueling
Praise
Lament
Group Singing

Individual Song Performance
Prayer
Sermon
Curing
Ceremonial Dialog

8. Element: Event

Name: Event

Identifier: Event

Definition: The event at which the content of the resource was obtained.

Comment: See list of Event Types, below.

Event Types

Interview
Linguistic Elicitation Session
Intra-Familial Social Gathering
Inter-Familial Social Gathering
Political Meeting
Religious Meeting
Community Gathering
Community Meal/Feast
Violent Confrontation
Storytelling Performance
Radio Program
Television Program
Ceremonial Greeting
Ceremonial Leave Taking
Public Announcement
Divination
Marriage Ceremony or Ritual
Birth Ceremony or Ritual
Puberty Ceremony or Ritual
Age-Grade Ceremony or Ritual
Funerary Ceremony or Ritual
Religious Ceremony or Ritual
Curing Ceremony or Ritual

9. Element: Transcription

Name: Transcription

Identifier: Transcription

Definition: If the resource is, or includes, a transcription of audio or video materials, the type(s) of transcription.

Comment: See list of Transcription Types below,

Transcription Types

Phonemic
Phonetic IPA
Phonetic Other
Practical Orthography
Indigenous Orthography
Prosodic Features
Conversation-Analytic
Musical
Gesture
Eye Gaze
Kinesthetics

10. Element: Translation

Name: Translation

Identifier: Translation

Definition: If this resource is, or includes, a translation, the type of translation.

Comment: See list of Translation Types, below

Translation Types

Morpheme-By-Morpheme
Parsing of Inflectional Categories
Word-By-Word
Calque
Sentence-Level Free Translation
Super-Sentential Free Translation
Interlinear