Customizing the IMDI metadata schema for endangered languages

Heidi Johnson
The Archive of the Indigenous
Languages of Latin America
The University of Texas at Austin
hjohnson@mail.utexas.edu
www.ailla.org

Arienne Dwyer
University of Kansas
DOBES (Volkswagen Foundation Programme
for the Documentation of Endangered Languages)
arienne@ku.edu

Abstract

The IMDI metadata schema must be adapted in order to accommodate endangered language data. The International Metadata Initiative is primarily oriented towards written standardized text, whereas most endangered languages are unwritten and nonstandardized. Archives and clearinghouses for endangered languages have a particular need to document unusual resources and types of metadata which may have no precedent in standardized written language. However, they must also support descriptions of written works, such as dictionaries, textbooks, and literature in endangered languages, which are produced through efforts to analyze, teach, or foster these languages. Our aim in subclassifying the content of endangered language data is to facilitate comparative research by developing a limited generic hierarchy with queryable controlled vocabularies of elements and sub-elements. Representing two endangered-language initiatives, the authors make specific recommendations for metadata categories, focusing on Content. We build on last years' DOBES recommendations to propose schemas generic and flexible enough to cover the range of resources likely to be housed in endangered language archives. The proposed schemas will be fully interchangeable with OLAC/DC metadata by means of conversion filters. Resources are archived in bundles of multimedia files with related content.

1. Introduction

Digital archives of resources in and about endangered languages have been established for the Volkswagen Foundation's Documentation of Endangered Languages initiative (DOBES) at the Max-Planck Institute for Psycholinguistics (MPI) in Nijmegen, and at the Archive of the Indigenous Languages of Latin America (AILLA) at the University of Texas at Austin. Both archives have adopted the International Standards for Language Engineering Metadata Initiative (IMDI) schema for describing the resources in their collections.

Earlier proposals for metadata for language resources have been oriented towards written standardized texts; however, most endangered languages do not have standard writing systems, and few enjoy a significant written literature as yet. Most of the resources collected thus far in both the DOBES and AILLA archives are audio and video recordings with accompanying textual annotations, such as transcriptions and translations. These recordings vary widely in terms of genre and communication context; it is the goal of both archival communities to record as much of the verbal art and discourse practices of the speakers of these languages as possible[1]. It is thus imperative that the metadata schema include categories for the description of many kinds of speech event and of many facets of the event.

However, not all of the resources in these archives are recordings or annotations of recordings: both archives include secondary resources such as grammars and dictionaries in their collections, and AILLA will archive pedagogical materials for language revitalization programs and written works created by native speakers of endangered languages as well. We must devise metadata schema that allow for a wide range of resources in and about these languages, recognizing that nearly any work may be either oral, written or gestural. (Word lists, for example, may be recorded, and interviews may be written.) The overall goal is to make the schemas generic and flexible enough to cover any theoretical bent, speech situation, or written form.

This paper presents a further refinement of the IMDI schema (ISLE group, 2001; Dwyer and Mosel, 2001) aimed at both facilitating the input of metadata and at facilitating searches of the archives' contents. We are chiefly concerned with definition of the Content subschema; however, section 2.0 contains a discussion of the general character of sets of multimedia language resources, and section 4.0 contains some discussion of the other subschema based on AILLA's customization experiences and in constructing a relational database based on the IMDI metadata elements.

2. Bundles of resources

The Metadata Elements for Session Descriptions (ISLE group, 2001) assembles the information for a group of related archive resources. The term *session* derives from a typical linguistic elicitation event: a recording of natural discourse performed by speakers of an endangered language, generally in the presence of a researcher who is engaged in an ongoing program of research. The session includes the digitized recording(s) and text files containing transcriptions, translations and other annotations.

This organization scheme ably captures the essential character of multimedia resources - that they come in sets of related files. However, not all sets of resources are based on recording sessions; a dictionary, for example, is typically a text file, which may have supporting audio files of examples, but can not be considered an annotation of those files. It is proposed that the general term *bundle* be used, to accommodate the full range of materials that will be housed in endangered language archives.

[1] Speech styles and genres are disappearing even more rapidly than the languages themselves, so there is urgent need to document ways of speaking as well as general facts about the language (Gnerre 1986).

A few examples of types of bundles may be useful in appreciated the subsequent discussion of the *Content* and other subschemas. The **canonical bundle** is the original session: digitized recordings with accompanying textual annotation files. Note that the media files may be archived in several formats (both .wav and .mp3 for audio; both .pdf and .doc for text, for example), and that a single session may have been recorded in multiple formats, so the bundle for a single session may consist of a dozen archived files.

A **minimal bundle** consists of a single file. A complete minimal bundle is most likely to be a text file, such as a dictionary or a poem, but it might also be a recording for which transcription and translation are not possible, such as a recording of chanting in a sacred unintelligible form of speech.

A **meta-bundle** is a bundle that contains other bundles. An example of this is a book such as Sherzer, 1990, which is an analysis of several oral performances, each of which is also archived in the form of canonical bundles, and may thus be searched for separately or included in other meta-bundles.

3. Overview of the IMDI schema

In the following discussion, names of schema categories or elements (topical labels) are presented in italics; the values of categories (provided by the metadata recorder) are underlined.

The schema consists of general facts about the bundle, such as the date and place of creation, and several subschemas:

1. ***Project*:** name and contact information for the larger context of the bundle.
2. ***Collector*:** name and contact information for the person who collected the bundle.
3. ***Content*:** a set of categories detailing the intellectual content of the bundle.
4. ***Participants*:** names, roles, and other information about the notable participants in the bundle-producing event.
5. ***Resources*:** information about the media files, such as URL, size, etc.
6. ***References*:** citations, URLs, etc. to relevant publications and other archive resources.

This design translates extremely well to other kinds of resource-producing events, including the writing of creative literature or compiling of a dictionary, since these activities are also often undertaken in the context of some project, and may involve a number of participants, such as authors, editors, translators, etc.

One addition that should be considered to the set of general *Bundle* elements is *Resource Relations*, which would define the interrelationships among items in the bundle. There are three kinds of relation that may obtain between two items in a bundle: derivation, sequence, or part-whole. Information about such relations is necessary so that users can reassemble the parts of a complex bundle correctly. At present, the value for this category will be a text description, separated from the general *Description* element for clarity. In the future, a more computationally tractable definition would be preferable.

Other simple addition are the element *Date Archived* and *Last Modified*, part of the general provenance of the bundle.

Lists of values for controlled vocabulary elements are given in the Appendix.

4. The Content Subschema

The controlling category for the content of resources for endangered languages is *Genre*, since the choice of genre may narrow the range of choices for other parameters, or even the set of additional parameters that must be defined. For example, if the resource is a dictionary, the *Modality* is writing and categories concerning the *Communication Context* are largely irrelevant. If the resource is a video recording of a ceremonial discourse, however, the *Modality* is speech (or gestural) and *Communication Context* categories are crucial pieces of information.

4.1. Genres

We propose a top-level division of the conceptual space of the *Genre* category into five subcategories:

1. ***Interaction*:** a discourse with two or more participants in which the central feature is the (verbal or written) exchange, i.e. conversation, argument, interview, etc.
2. ***Explanation*:** discursive (typically monologic) genres, i.e. statement, description, procedure, etc.
3. ***Performance*:** an audience is the central feature of performances; however, in elicitation settings or for written works the audience may be more virtual than actual. Examples: narrative, oratory, poetry, etc.
4. ***Teaching*:** pedagogical materials not included in another genre, i.e. primer, textbook, etc.
5. ***Analysis*:** the products of scholarly research, i.e. grammar, dictionary, sketch, etc.

Any given resource may in fact participate in several of these genres (a *Performance* may include *Interaction*, for example), but it is impossible to perfectly decompose the range of human expression into discrete categories. The metadata recorder is expected to choose based on the predominant characteristics of the resource in question, and to provide additional information in the *Description* section of the *Content* schema.

4.2. Other Content categories

The other *Content* categories in the current IMDI schema are:

1. ***Modality***: speech, writing, gesture.
2. ***Communication Context*:** a sub-subschema with three parts:
    a. *Interactivity*: interactive, non-interactive, semi-interactive.
    b. *Planning*: spontaneous, semi-spontaneous, planned.
    c. *Involvement*: unmarked, elicited, non-elicited, observer-absent.
3. ***Languages*:** all the language varieties used in the bundle.

4. ***Task*:** a specific research task or experiment, such as "info kiosk task".
5. ***Description*:** text that adds details about the bundle.
6. ***Keys*:** customizable Key-Value pairs that may be added by particular organizations.

The AILLA team has added the following categories as Key-Value pairs, which may be considered for adoption in the general IMDI schema:

1. ***Register:*** a rough characterization of the way in which the discourse reflects the discourse situation, i.e., informal, formal, honorific, etc.
2. ***Style:*** a broad category intended to capture a range of poetic and stylistic effects, such as metered lines, play language, parallelism, etc.

Other information, such as details about the setting in which the resource was produced, are best described in prose passages in the *Description* element.

5. The Non-Content Schemas

In this section we review the remaining subschemas in light of customizations that AILLA has made for its holdings, to demonstrate the flexibility of the current IMDI design and, simultaneously, suggest elements that might be incorporated into the standard schema. All of the subschemas provide a *Description* element for further information. Only the *Participant* subschema allows customizable sets of Key-Value pairs. We suggest that these be made available for all subschemas, to allow maximum flexibility for archives using the IMDI schema.

5.1. Project

The Project subschema consists of the following elements:

1. ***Name:*** an abbreviated name for the project.
2. ***Title:*** the full name of the project.
3. ***Id:*** a unique identifier.
4. ***Contact:*** a subschema for address, email, telephone, and other contact information.

AILLA adds a *Funder* element whose value is the name of the organization that funded the project. The *ID* element is not used.

One rather trivial comment here is the potential for confusion that arises from using the term 'Name' to refer to an acronym, initials, or nickname (see 5.3, *Participants*). We would suggest calling this element *Short Name*, and placing it after *Title* (or *Full Name* element for *Participants*).

5.2. Collector

AILLA renamed this subschema *Depositor*, since this is the individual for whom contact information must be maintained, and because in most cases, the Collector and the Depositor are the same person. In cases in which they are different people, the Collector can be identified in the *Participants* group.

5.3. Participants

The IMDI schema collects considerable information about the persons who were notably involved in the production of the resource. This is important for a true understanding of the character and quality of the data, and also to ensure that credit is given where credit is due. The categories in the Participants subschema are these:

1. ***Type:*** the functional role of the participant, i.e. creator, interviewer, translator, etc.
2. ***Role:*** intended to refer to family or other personal relationships amongst participants.
3. ***Name/Full Name:*** a short form, such as an alias or initials, and the full name.
4. ***Language:*** languages spoken by the participant, listing the native language first.
5. ***Ethnic group, Age, Sex*:** self-explanatory.
6. ***Education*:** highest level of education attained.
7. ***Anonymous:*** True if the participant's Full Name is reserved; False otherwise.

AILLA doesn't use the *Role* element, but has added two Key-Value pairs:

1. ***Origin:*** the place of origin of a participant, of interest in sorting out dialect issues. This element would be best coded as having a subschema value (see 5.5).
2. ***Occupation:*** the principal occupation of the participant. This can be an important piece of information about the particular resource (for example, a healing chants sung by a healer may be considered more authoritative than one sung by a non-specialist), but is also an interesting part of the participant's history that merits recording.

5.4. Resources

The first suggestion is to change the name *Annotation File* to *Text File*, since text files are not necessarily annotations of some recording. Minimal bundles are most likely to be single text files, such as dictionaries or grammatical sketches.

The elements that describe *Text Files* also need some clarification and expansion. Currently, there are three:

1. ***Type:*** type of annotation, e.g., phonetic transcription.
2. ***Content Encoding:*** the annotation encoding scheme, e.g. EUROTYP morphosyntactic annotations.
3. ***Character encoding:*** the character set(s) used in the text.

The term *Type* is confusing in this context. We suggest the phrase *Transcription Type*, which although cumbersome, has the advantage of being unambiguous. AILLA adds a parallel element *Glossing Type* (e.g., morpheme-by-morpheme), and *Software* (e.g., Shoebox) for identifying software used in transcription or translation. Information about the person who produced

the transcription or translation (currently encoded in the element *Annotator* in the *Text File* subschema), should be given in the *Participants* group, which would allow information about the annotator's languages to be encoded; an important datum in considering translations.

Finally, the two top-level categories, *Media File* and *Annotation File*, have a number of elements in common, which is problematical for the construction of a relational database, because it introduces significant redundancies.

A more database-friendly design would list common elements directly under the Resources heading, and separate only those elements that are unique to each type of resource into separate subschemas. A general *Type* element would be useful here, with values audio, video, image, and text. This design will be easier to extend, should we encounter additional types of resources.

The resulting Resources subschema looks like this:

Resources	
Resource link	URL = Identifier
Type	{audio, text, etc.}
Size	in bytes
Format	MIME
Access	permissions subschema
Language	
Anonymous	
Media File	
Quality	rating from 1-5 (low-high)
Recording conditions	describe equipment, etc.
Position	if a segment of larger file
Text File	
Content encoding	
Character encoding	
Transcription type	
Translation type	
Software	

5.5. Other subschemas

Several smaller schemas are defined for structured information that occurs repeatedly. One example is the *Description* element, which is a short descriptive text with a code identifying the language used. The other subschemas are *Access*, a set of elements for specifying the access permissions associated with the bundle; *Keys*, a structure for defining custom elements; and *Language*, which includes elements for the language code (Aristar & Dry, 2001), the variant *Names* of the language, and a *Description* element for additional information; and *Contact*, which collects contact information for a person or organization.

We suggest that a new subschema *Place* would be useful, since this is also a structured element that appears more than once (place where the bundle was produced; place of *Origin* of a *Participant*.) This subschema would have the following elements: *Continent*, *Country*, *Region*, *SubRegion* (which could hold an actual address).

6. Conclusion

The IMDI schema is a flexible tool for characterizing a wide variety of speech genres, with customization facilities encoded in the optional Keys subschema. The proposals here will extend the schema to encompass textual resources as well as oral works, and facilitate the metadata definition task by grouping related elements in coherent subsets. It also readily supports implementation of a relational database. It should also be noted that conversion filters can be written to map from the IMDI schema to the OLAC/DC metadata, ensuring compatibility across these major sets of metadata for endangered languages.

7. References

Aristar, Anthony and Helen Dry. (2001). *The EMELD Project.* Proceedings of the IRCS Workshop on Linguistic Databases, Philadelphia, 11-13 December 2001. pp. 11-16.

Dwyer, Arienne and Ulrike Mosel. (2001). *Metadata Description Recommendations: General; Content.* DOBES Technical Reports 6a.1, 6b.2 (15.03.01).

Gnerre, Maurizio. (1986). *The decline of dialogue: Ceremonial and mythological discourse among the Shuar and Achuar of Eastern Ecuador.* In Joel Sherzer and Greg Urban Eds., Native South American discourse. Berlin: Mouton de Gruyter. pp. 307-341.

ISLE Group. (2001). *Metadata Elements for Session Descriptions.* ISLE Metadata Initiative, Draft proposal version 2.4 (7), May, 2001. http://www.mpi.nl/ISLE/documents/docs_frame.html

Sherzer, Joel. (1990). Verbal Art in San Blas: Kuna culture through its discourse. Cambridge: Cambridge University Press.

Simons, Gary and Steven Bird. (2001). *OLAC Metadata Set.* Draft, Open Language Archive Consortium.

8. Appendix: Controlled vocabularies

Genre:

- ***Interaction:*** conversation, verbal contest, interview, meeting/gathering, riddling, consultation, greeting/leave-taking, humor, insult/praise, letter.
- ***Explanation:*** procedure, recipe, description, instruction, commentary, essay, report/news.
- ***Performance:*** narrative, oratory, ceremony, poetry, song, drama, prayer, lament, joke.
- ***Teaching:*** textbook, primer, workbook, reader, exam, guide, problem set.
- ***Analysis:*** dictionary, word-list, grammar, sketch, field notes.

Register: informal/conversational, formal, honorific, jargon, baby/caretaker talk, joking, foreigner talk.

Style: ordinary speech, code-switching, play language, metrical organization, parallelism, rhyming, nonsense/unintelligible speech.

Participant:

Type: narrator, interviewer, respondent, author, photographer, filmer, consultant, researcher, transcriber, translator, annotator, recorder, editor, interlocutor.

Transcription Type: phonemic, phonetic IPA, phonetic other, practical orthography, prosodic features, conversation-analytic, musical, gestural, eye-gaze, kinesthetic.

Translation Type: morpheme-by-morpheme, word-by-word, interlinear, sentence-level free translation, super-sentential free translation.