tSouthern Oregon University
Department of Computer Science
Technical Report CS0809-01, May 23, 2010

ACORNS Multilingual Dictionary XML (AMDX)

Lewis Nakao and Dr. Dan Harvey

Abstract

The ACORNS project, acronym for [A][C]quisition [O]f [R]estored [N]ative [S]peech, develops software tools to support indigenous tribal efforts to preserve and revitalize their languages and cultures. The name, ACORNS, is to honor the tribes of Northern California, who hold the acorn sacred and who, in 2005, helped to spawn this long-term development effort. The software is intuitive and easy to use, without requiring significant technical training. The ACORNS project now provides seven categories of lessons that language teachers or their students can prepare to create a library, which can be made available either on the Web or through CD distribution.

Recently, we embarked to design a dictionary-based application, WOLF ([W]ord [O]riented [L]inguistic [F]ramework), that linguists can use for language archival, but also to provide data for dictionary-based games and game engines that employ automatic speech recognition (ASR). This report documents the XML format for files imported to and exported from WOLF.

The design goals for WOLF present many difficult challenges. For example, linguists employ a wide variety of tools to create their dictionaries. These range from non-linguistic applications like EXCEL, to those designed for dictionary data like Toolbox, Lexique Pro, or We Say. An effective design should be general enough to be able to import data from various formats, and export the data to those same formats.

The structure needs to be XML-based and well-documented if it is to follow the best-practice standards established by Electronic Metastructure for Endangered Languages Data (EMELD). These standards resulted from a five-year project and they are published at http://emeld.org/index.cfm. We must also maintain the central requirement for all ACORNS development; the software must be language independent so it will be equally effective within the context of any language.

This paper describes a first version of the ACORNS Multilingual Dictionary XML (AMDX) format. After researching various existing formats, data storage standards, and linguistic software, we feel that this standard satisfies the requirements that we describe above. This report is meant to be a starting point for ongoing development. Involvement and collaboration from the linguistic community is necessary, and this process will provide feedback and discussion that will lead to further refinements.

General Characteristics of the AMDX Format

The initial requirement that was considered in developing the AMDX format was flexibility in format structure, which is lacking in many dictionary formats. Often a particular format finds a limited user base because: its design is tailored to the requirements of a particular language; it uses a proprietary internal format; or adequate documentation is lacking. Our design stresses flexibility and incorporates the following features:

1. AMDX format enables a single dictionary to hold data for multiple languages. This feature enables linguists to reference words or phrases of different languages within a single source and this can lead to exciting opportunities for linguistic research.
2. *The Ethnologue*, an SIL International sponsored resource, distributes a catalog of three-letter language codes that uniformly identifies all known languages with their alternate names (also known as ISO/DIS 639-3). The Ethnologue and the ISO worked in cooperation to create this international standard. The latest version, the 15th edition, was released in 2005 and is available in downloadable database tables. (Gordon, 2005). The AMDX format uses these three-letter language codes in many of its attributes to identify languages. For a reference to these language codes and downloadable tables, please visit the Ethnologue website (http://www.ethnologue.com/codes/).
3. The WOLF dictionary application makes use of drop-down menus to accommodate the GOLD ontology. It also enables linguists to customize this ontology to their individual needs, if necessary. This approach achieves maximum flexibility; yet encourage users to utilize standard terminology, which enables data to be easily indexed and uniform.
4. The font support facilities allow the use of any custom indigenous font installed on a particular system. This capability enables use of the International Phonetic Alphabet

(IPA) to represent the phonetics that make up words and phrases. WOLF also supports, on all platforms, keyboard mappings created in *.keylayout* format, popular on MAC systems.

5. The AMDX format allows multimedia files to be attached to dictionary-data using allowable syntax. For example, audio, video, or still-pictures can be associated with a word, a definition, or an example. To conform to best-practice principles, these files, when exported, are to be stored in well-known formats (e.g. jpg, wav, mp3, etc.).
6. The format does not specifically address the issue of handling legacy data. However, applications that utilize AMDX dictionary documents should provide import and export facilities to enable linguists to transport data back and forth between AMDX-based applications and other applications that they might prefer.

The features of AMDX are likely to be more than most linguists will employ when working with a single language. Our goal is for the WOLF dictionary application to provide an easy to use interface that will enable users to easily build dictionaries using a subset of these features.

AMDX General Structure

The following figure briefly presents the basic elements of an XML document that will contain dictionary data. For brevity, there are a couple of elements that are not listed in this list; however, all element and sub-elements are explained in detail in subsequent sections of this report.

- Top level: <amdx>
  - Author list: <authors> (<author> list)
  - Copyright: <copyright>
  - Language: <languages> (<language> list)
    - Word list: <words> (<word> list)
      - Define a word: <word>
        - Multimedia: <media>
        - Language Translations: <translations> (<translation> list)
        - Columns: <columns> (<translations>, <classification>, <example>, <ontology> list)
        - Rows: <rows> (<example>, <classification>, and <definition> list)
          - Example: <example>
            - Multimedia: <media>
            - Language Translations: <translations> (<translation> list)
          - Comment: <classification>
          - Definition: <definition>
            - Multimedia: <media>
            - Language Translations: <translations> (<translation> list)
            - Columns: <columns> (<classification>, <example>, and<ontology>, <translations> list)
            - Rows: <rows> (<example> and <classification> list)
              - Example: <example>
                - Multimedia: <media>
                - Language Translations: <translations> (<translation> list)
                - Comment: <classification>

Figure 1 The AMDX format structure overview.

AMDX DTD Definition

The AMDX format must follow the XML DTD (Document Type Definition) below:

```
<?xml version="1.0"?>
<!DOCTYPE amdx SYSTEM "amdx.dtd">

<!-- <amdx> root element -->
<!ELEMENT amdx (authors?, copyright?, languages?)>
   <!ATTLIST amdx version CDATA #REQUIRED >
   <!ATTLIST amdx created CDATA #IMPLIED >
   <!ATTLIST amdx modified CDATA #IMPLIED >
   <!ATTLIST amdx face CDATA #IMPLIED >
   <!ATTLIST amdx size CDATA "12" >

<!ELEMENT copyright (#PCDATA) >
   <!ATTLIST copyright date CDATA #IMPLIED >

<!-- <author> element -->
<!ELEMENT authors (author*) >
<!ELEMENT author EMPTY >
  <!ATTLIST author name CDATA #IMPLIED >
  <!ATTLIST author org CDATA #IMPLIED >
  <!ATTLIST author email CDATA #IMPLIED >
  <!ATTLIST author url CDATA #IMPLIED >
  <!ATTLIST author initials CDATA #IMPLIED >
  <!ATTLIST author langs CDATA #IMPLIED >

<!-- <language> element -->
<!ELEMENT languages (language*) >
<!ELEMENT language (words?) >
   <!ATTLIST language lang ID #REQUIRED >
   <!ATTLIST language variant CDATA #IMPLIED >
   <!ATTLIST language sort CDATA #IMPLIED >
   <!ATTLIST language face CDATA #IMPLIED >
   <!ATTLIST language size CDATA "12" >
   <!ATTLIST language name CData #IMPLIED >

<!—elements for a list of words -->
<!ELEMENT words (word*) >
<!ELEMENT word (media?, translations?, columns, rows) >
   <!ATTLIST word width CDATA #IMPLIED >

<!-- <definition> element -->
<!ELEMENT definition (media?, translations?, columns, rows) >
   <!ATTLIST definition width CDATA #IMPLIED >


<!-- <example> element -->
<!ELEMENT example (media?, translations?) >
   <!ATTLIST example title CDATA #IMPLIED >
   <!ATTLIST example width CDATA #IMPLIED >
```

```
<!–elements for a multimedia (used in multiple instances) -->
<!ELEMENT media EMPTY >
   <!ATTLIST media audio CDATA #IMPLIED >
   <!ATTLIST media video CDATA #IMPLIED >
   <!ATTLIST media picture CDATA #IMPLIED >

<!-- <translation> element (used in multiple instances) -->
<!ELEMENT translations (translation*, #PCDATA) >
   <!ATTLIST translations phonetics CDATA #IMPLIED >
   <!ATTLIST translations title CDATA #IMPLIED >
   <!ATTLIST translations width CDATA #IMPLIED >
<!ELEMENT translation (#PCDATA) >
   <!ATTLIST translation lang IDREF #REQUIRED >


<!-- <columns element (used in multiple instances) -->
<!ELEMENT columns (classification*, ontology*, translations*) >

<!-- <rows element (used in multiple instances) -->
<!ELEMENT rows (definition*, example*, classification* ) >

<!-- <classification> element (used in multiple instances) -->
<!ELEMENT classification (#PCDATA) >
   <!ATTLIST classification face CDATA #IMPLIED >
   <!ATTLIST classification size CDATA "12" >
   <!ATTLIST classification width CDATA #IMPLIED >
   <!ATTLIST classification phonetics CDATA #IMPLIED >
   <!ATTLIST classification title CDATA #IMPLIED >

<!-- <ontology> element (used in multiple instances) -->
<!ELEMENT ontology (#PCDATA) >
   <!ATTLIST ontology parent CDATA #IMPLIED >
   <!ATTLIST ontology child CDATA #IMPLIED >
   <!ATTLIST ontology abbreviation CDATA #IMPLIED >
   <!ATTLIST ontology phonetics CDATA #IMPLIED >
   <!ATTLIST ontology type CDATA (0|1|2|3|4) >
   <!ATTLIST ontology width CDATA #IMPLIED >
```

XML Structure

This section describes the details of AMDX syntax. In the descriptions of the XML elements, *[text]* refers to the character content between the open and close elements (e.g. *<xmltag>* the value in here *</xmltag>)*. In general, the description of the format will begin by describing the root element and traversing downward to the sub-elements. However, some elements that are used in multiple places will only be described once. There are examples given, but please note;

these are just for illustration; improvements are possible.

<*amdx*> (required)

The top level element, <*amdx*>, is required. It has five attributes, of which *version* is required, and *created* and *modified* attributes are optional should the dictionary creator want to enter the corresponding dates. The *face* attribute specifies the font used to display phonetic symbols in the dictionary and the size attribute specifies the size of that font. There are three sub-elements, <*copyright*> and <*authors*> and <*languages*>. The <*languages*> sub-element specifies characteristics that are unique to a particular language and includes the words associated with that language. The optional <*copyright*> sub-element contains copyright text and an <*authors*> sub-element credits contributors to the dictionary. Subsequent sections of this report provide more details relating to each of these sub-elements.

version	(required) Indicates the version of AMDX dictionary format being used.
created	(optional) Indicates the date and time when the dictionary was created.
modified	(optional) Indicates the date and time when dictionary was last modified.
face	(optional) Indicates the name of the font used to display phonetic symbols.
size	(optional) Indicates the size of the font used to display phonetic symbols

Table 1 Listed attributes of <amdx> elements

Example:

```
<amdx created="2007-11-21" modified="2008-12-31" version="1.05" face="ipasamm" >
```

<*copyright*>

The <*copyright*> element is optional and does not have any sub-elements. It has one attribute.

[text]	(optional) Contains any use restrictions associated with a dictionary.
date	(optional) The year originating the copyright.

Table 2 Listed attributes of <copyright> elements

Example:

```
<copyright date="2009-04-30">Contact Lewis Nakao for permissible use</copyright>
```

<authors>

The *<authors>* element contains a list of *<author>* sub-elements and has no attributes.

<author>

The *<author>* element has no sub-elements and contains the following attributes that describe a contributor to the dictionary.

name	(optional) Contains the name of the author who has contributed to the dictionary.
org	(optional) Contains the name of the organization the author represents.
email	(optional) Contains the email of the author.
url	(optional) Contains the website of the organization or author.
initials	(optional) Contains the initials used to reference dictionary authors; each unique to that of other authors.
langs	(optional) Contains a list of three-lettered ISO/DIS 639-3 language codes, separated by commas, that indicates what languages the author was responsible for in this dictionary. The languages are separated by commas (if more than one).

Table 3 Listed attributes of <author> elements

Example:

```
<authors>
   <author name="Lewis Nakao" org="Southern Oregon University"
         email="lewdev@lewdev.net" initials="LTN" langs="eng,jpn" />
   <author name="Dan Harvey" org="Southern Oregon University"
         url="http://cs.sou.edu/~harveyd" email="harveyd@sou.edu"
         initials="DH" langs="eng" />
</authors>
```

<languages>

The *<languages>* element contains a list of *<language>* sub-elements; it has no attributes.

<language>

The *<language>* element optionally contains a single *<words>* sub-element, which contains the list of words in the language. It also contains four attributes that describe characteristics specific to the language. These attributes are listed in the table below.

name	(optional) Contains the name of the language. The default is the name found by looking up the ISO/DIS-639-3 language code table.
lang	(required) Contains the language to be used in the dictionary using the three-letter ISO/DIS 639-3 language code. If there is a dialect variant, one or two alphabetic characters follows the language code with a / delimiter (*lang/variant*). The variant code must match the value of the variant attribute (below). There cannot be two <language> elements with the same lang value.
variant	(optional) One or two alphabetic characters to specify a particular writing system or alternate dialect.
sort	(optional) Custom sort order containing sequence of characters. The earlier characters are first in the collating sequence.
face	(optional) Indicates the font used for the language, which is assumed to be installed on the user's computer or supported by the application.
size	(optional) The size of font when used in an application. By default the font size is 12.

Table 4 Listed attributes of <language> elements

Example:

```
<languages>
   <language id="eng" face="Verdana" size="12" name=”English” >
   <words>
         <!–list of words go here –>
   <word>
   </language>
   <language id="jpn" face="JP-Font" size="10" sort="DdCcBbAa"
   name="Japanese" >
   <words>
         <!–list of words go here –>
   <word>
   </language>

</languages>
```

<*words*> (required)

The <*words*> element contains a list of <*word*> sub-elements without any attributes.

<word>

The <*word*> element fully describes a particular word in a dictionary. This element reflects the WOLF application interface where word-oriented information is entered into row and column cells. This concept is similar to that of spread sheet programs like EXCEL, but the wolf interface is geared for use by linguists. The <*word*> element includes <*rows*> and <*columns*> sub-element for specifying cells, each of which contain word-based data. It has a single attribute that specifies the width of the cell for entering the word.

A <*word*> element defines a <*media*> sub-element for attaching audio, pictures, and video. It also specifies a <*translations*> element enabling a dictionary to translate words into other languages.

width	(optional) Contains the width of the cell in pixels for user entry of word spellings and their translations.

Table 5 Listed attributes of <word> elements

Example:

```
<words>
   <word width="200">
      <media picture="greeting.png" />
      <translations>Hello
         <translation lang="jpn">今日は</translation>
      <translations>
      <columns><!—word-based data goes here —></columns>
      <rows><!—word-based data goes here —></rows>
   </word>
</words>
```

<definition>

The structure of <*definition*> elements is identical to that of <*word*> elements. It also can contain <*media*> and <*translation*> sub-elements and information is entered in rows and

columns (represented by the <*rows*> and <*columns*>) sub-elements. There are a couple of details that differ, however. The first is that it is illegal to add a definition to a definition, since this would not make sense. Definitions also allow a <*translations*> column, which specifies words in other languages having the same definition. This capability is not needed when a word is specified because words can already link to translations in other languages. Definition elements can contain a *width* attribute to define the width, in pixels, of a definition entry. Please refer to the description of the <*rows*> and <*columns*> elements for more details.

width	(optional) Contains the width of the cell in pixels for user entry of a definition and its translations.

Table 6 Listed attributes of <word> elements

Table 9 The <definition> data

Example:

```
<definition>
   <media picture="greeting.png" />
   <translations>Used to express a greeting or answer a telephone
     <translation
     lang="jpn">あいさつを表明することには電話、または使用されている答えの注目を集める。
     </translation>
   </translations>
   <columns><!—definition-based data goes here —></columns>
   <rows><!—definition-based data goes here —></rows>
</definition>
```

<examples>

The <*examples*> element contains of a list of <*example*> sub-elements; it has no attributes.

<example>

The <*example*> element has a similar structure as <*word*> and <*definition*> elements. The major difference is that <*example*> elements do not contain <*rows*> and <*column*> sub-elements. It does allow a <*media*> sub-element for attaching audio, video and pictures and <translations> element enabling the example to be translated into other languages. It has one attribute, *width*, defining the width of the component in pixels.

title	(optional) The type of example. For example, a Lexical Function
width	(optional) Contains the width of the cell in pixels for user entry of example text and their translations.

Table 7 Listed attributes of <example> elements

Example:

```
<example>
   <translations> After their hellos, they engaged in deep conversation
         <media picture="conversation.jpg" />
         <translation lang="jpn">の挨拶を経て、深い会話に従事。</translation>
   </translations>
</example>
```

<media>

The *<media>* element is used in various sections of an AMDX document to include multi-media elements. This element has three attributes that respectively define locations for audio, video, and picture resources. The initial WOLF application expects the actual source for the resource to be contained in a subfolder by the same name of the dictionary, but without the *xml* extension. In future versions, we plan to also allow URL locations.

audio	(optional) name of an audio resource.
video	(optional) name of a video resource.
picture	(optional) name of a picture resource.

Table 8 Listed attributes of <media> elements

Example:

```
<media audio="cat.wav" video="cat.mov" picture="cat.jpg" />
```

<translations>

The *<translations>* element contains gloss and phonetic text for a word, definition, example, commnet, or column elements. It also defines a list of *<translation>* sub-elements to translate the gloss text into other languages.

[text]	(optional) Contains the gloss text for a word, definition, or example.
phonetics	(optional) Phonetic representation of the gloss text for a word, definition, or example.
title	(optional) Type of column element. Values include Compare, Encyclopedic Info, Gloss, References, Restrictions, Usage, and Variants. The default value is References.
width	(optional) Contains the width of the cell in pixels for user entry of reference text and their translations.

Table 9 Listed attributes of <translation> elements

<translation>

The *<translation>* element requires a *lang* attribute and contains no sub-elements. The text content is a direct translation of the word, definition, or example specified in a parent element. It applies to the language corresponding to its *lang* attribute value.

[text]	(optional) Contains the translation of a word definition or example using the language indicated in the *lang* attribute.
lang	(required) Contains a three-lettered ISO/DIS 639-3 language code that determines what language *[text]* is using.

Table 10 Listed attributes of <translation> elements

Example:

```
<translations>hello
   <translation lang="jpn">今日は</translation></translation>
</translations>
```

<columns>

The *<columns>* element specifies word and definition oriented data that applications should present as column cells. The sub-elements include *<classification>*, *<ontology>*, and *<translations>*, each described in separate sections of this report. Of these, the *<translations>* element is only valid when the parent element is a definition.

<rows>

The *<rows>* element specifies word and definition oriented data that applications should present

as row cells. The sub-elements include <*classification*>, <*definition*>, and <*example*>. The last two of these was described above, and the <*classification*> element's description is forthcoming. Note that it is illegal to include a <*definition*> element as a row of a definition.

<classification>

The <*classification*> element specifies text to be added to a word or definition. If it has a title attribute, it identifies the type of entry (i.e Synonyms, Antonyms, etc.). The WOLF application provides default classification titles. These are *Antonyms, Categories, Etymology, Language Links, Main Entry, Refer To, Reversals, Spellings, Subentry, Synonyms, Table, and Thesaurus*. It anticipates, but doesn't require, the text for *Antonyms, Categories, Main entry,* and *Thesaurus* to be a series of words separated by commas, spaces, or semicolons. Categories are a classification (like animal or mammal). It uses this format to perform category searches of the dictionary. When there is no title attribute, the data is free-form text that can be any comment or culturally relevant information.

The other attributes allow specification of phonetics, cell width in pixels, and a font to use with the <*classification*> cell in question.

title	(optional) Identify the type of classification. Values include: Antonyms, Categories, Etymology, Language Links, Main Entry, Refer To, Reversals, Spellings, Subentry, Synonyms, Table, and Thesaurus.
phonetics	(optional) Phonetic representation of the data entered.
width	(optional) Contains the width of the cell in pixels for user entry of text.
face	(optional) Indicates the font to use for this cell.
size	(optional) The size of font to be used in this cell. By default the font size is 12.

Table 11 Listed attributes of <language> elements

Example:

<classification face="aboriginal sans" size="10" title="Synonyms">hi</classification>

<ontology>

The <*ontology*> element relies on GOLD ontology specifications for ontological elements. It has no sub-elements, but contains a number of attributes. We specify each of these below:

parent	(optional) Specifies the ontological term in the GOLD ontology.
abbreviation	(optional) The abbreviation the dictionary creator prefers to use for the ontological term.
child	(optional) The value for the parent ontological term. If the parent attribute exists, so must the child.
phonetics	(optional) If the word form is present, this attribute enables specification of its phonetic representation.
type	(optional) There are five ways that the WOLF application can format an ontological cell. This integer specifies this formatting. Type 0 shows only the child value; type 1 shows parent and child on one line; type 2 shows parent and child on two lines; type 3 shows child on one line and allows user input of the word-form below; type 4 shows parent and child on one line and allows user input of the word-form
width	(optional) Contains the width of the ontological cell in pixels.

Table 12 Listed attributes of <ontology> elements

Example:

```
   <ontology child="Past" parent="Tense"
             phonetics="" type="3" width="108">threw</ontology>
```

Formatted Example

This section presents a limited example of the AMDX format that can serve as a reference that illustrates correct AMDX syntax. There are two languages used in this example, Japanese and English, and the word, “hello.” This sample provides a simple example; it is in no way comprehensive.

```
<amdx created="2007-11-21" modified="2008-12-31" version="1.05" face="ipa samm" >
<copyright date="2009-04-30">Contact Lewis Nakao for permissible use</copyright>
<authors>
   <author name="Lewis Nakao" org="Southern Oregon University"
           url="http://lewdev.net" email="lewdev@lewdev.net" initials="LTN"
           langs="eng,jpn"></author>
   <author name="Dan Harvey" org="Southern Oregon University"
           url="http://cs.sou.edu/~harveyd" email="harveyd@sou.edu"
           initials="DH" langs="eng">Dan Harvey</author>
</authors>
```

```
<languages>
   <language name="English" code="eng" font="Verdana" size="12" gloss="true">
   <words>
      <word>
      <media <media audio=="dictionary/hello.wav" />
      <translations phonetics="hɛloʊ, hə-,hɛloʊ">hello
            <translation lang="jpn">今日は</translation>
      <translations>
      <columns>
         <ontology abbreviation="intj" child="Interjection" parent="Part Of Speech"
                   type="2" width="110"/>
         <classification title="Synonyms"><synonym>hi</classification>
         <classification title="Antonyms">bye, depart</classification>
      </columns>
      <rows>
         <definition>to express a greeting or answer a telephone
         <translations>
            <translation lang=”jpn”>
               あいさつを表明することには電話、または使用されている答えの注目を集める。
            </translation>
         </translations>
         <example>
            <translations> Hello! It is good to see you!
            <translation lang=”jpn”>こんにちは！またお会いできて良いです！</translation>
            </translations>
         </example>
         </definition>
      </rows>
      </word>
   </words>
   </language>
   <language name="Japanese" code="jpn" font="JP-Font" size="10" ></language>
</languages>
</amdx>
```