HIS Registry of Dialects

Abstract: Documents the Registry of Dialects (ROD) of the Harvest Information System (HIS). This registry identifies dialects and defines a standardized code for each dialect.

Editor: Gerry Gutierrez, GRN Gerrygutierrez@globalrecordings.net

Table of contents

1. Overview

The function of the Registry of Dialects (ROD) is to (a) Identify specific varieties of given languages (defined by ISO 639-3) that research has determined to require distinct presentations (such as audio, video or print) in order to overcome barriers of understanding or acceptance. Determining factors may include differences in vocabulary, grammatical construction, idioms, and marked accents, as well as religious or social barriers. (b) Provide unique, standardized codes for these dialects

The registry contains a code table, a supplementary table and a change history table:

- ROD_Dialect: Code for Dialect

  A code in this code set represents a unique dialect of a language, as determined by the demonstrated need for distinct media presentations designed to overcome barriers of understanding or acceptance.

- ROD_AlternateNameIndex

  This table provides an index into ROD_Dialect based on alternate dialect names.

- ROD_DialectChangeHistory

  This table documents changes to the code set for dialects.

The code table and the supplementary table make use of the ROL_Language code table from the HIS Registry of Languages.

2. Code tables

The registry contains one code table:

ROD_Dialect: Code for Dialect

The code table contains a set of dialects of living languages that have been determined to require distinct media presentations.

By definition the scope of a dialect code is always a smaller group of speakers than the group represented by the assigned language as a whole.

Each code is a standardized five-digit code (including leading zeros, when necessary) for uniquely referring to a particular dialect. Using a code from this set not only uniquely identifies the dialect, but also identifies the language of which it is a part, through its corresponding ISO 639-3 (ROL) language code in the code table entry.

Additional information on the dialect may also be available in the *Global Recordings Network* website. Users have full access to those descriptions as follows:

(a) For any ROD dialect code *ddddd,* the following URL lists available information:

http://globalrecordings.net/dialect/ddddd

For example: **http://globalrecordings.net/dialect/04231** will show information for ASMAT: Waganu (ROD dialect #04231) including links to other sites.

(b) For any ROL language code xxx the following URL lists information on related dialects:

http://globalrecordings.net/langcode/xxx

For example: http://globalrecordings.net/langcode/asc will show information about dialects of the ASMAT language

The code table for ROD_Dialect contains the following columns:

Column	Format	Description
Dialect Code	char(5)	The five-digit numerical code for the Dialect.
ISO	char(3)	A three-letter code from the ROL_Language (ISO 639-3) code set that identifies the Language associated with the Dialect.
ROG3 Code	char(2)	A two-digit code from the ROG_Political code set that identifies the Country. These codes correspond with the FIPS2 codes
Dialect Name	varchar(75)	The primary name of the Dialect.
Language Name	varchar(75)	The primary name of the Language
Location	varchar(75)	The name of the country where the dialect is spoken. This may be followed by the province and/or district.

The SQL statement for creating this table is as follows:

```
CREATE TABLE ROD_Dialect (
    Code     char(5) NOT NULL,
    ISO      char(3) NOT NULL,
    ROG3     char(2) NOT NULL,
    Name     varchar(75) NOT NULL,
    Language varchar(75) NOT NULL,
    Location varchar(75) NOT NULL)
```

For example, typical entries for dialects of the Southern Pashtu language will look like this:

Code	ISO	ROG3	Dialect Name	Language Name	Location
00090	pbt	AF	Kabuli	Pashto, Southern	Afghanistan
00843	pbt	AF	Baluchi	Pashto, Southern	Afghanistan
00853	pbt	AF	Western	Pashto, Southern	Afghanistan
15545	pbt	AF	Kandahar Pashto	Pashto, Southern	Afghanistan
15546	pbt	AF	Qandahar Pashto	Pashto, Southern	Afghanistan
15547	pbt	AF	Quetta Pashto	Pashto, Southern	Pakistan
15548	pbt	AF	Southeastern Pashto	Pashto, Southern	Pakistan
15549	pbt	AF	Southwestern Pashto	Pashto, Southern	Afghanistan

Because dialects are linked to languages, many of their names (e.g. "West") do not stand alone. In the diagram above, dialect code 00853 represents the Western dialect (or variety) of the Southern Pashtu language in Afghanistan. Language names are quoted directly from the Ethnologue. Where possible, dialect names indicate the preference of the speakers.

3. Other tables

The registry contains one supplementary table.

ROD_AlternateNameIndex

This supplementary table offers an index into the dialect codes table by means of alternate names. Whereas the ROD_Dialect table lists only primary dialect names, this index makes it possible to find a code by any one of the alternate names for a given dialect.

The ROD_AlternateNameIndex contains the following columns:

Column	Format	Description
Dialect Code	char(5)	The five-digit numerical code for the dialect from ROD_Dialect
Variant	Varchar(75)	A name associated with the dialect

```
CREATE TABLE ROD_AlternateNameIndex (
    Code        char(5) NOT NULL,
```

```
Variant varchar(75) NOT NULL)
```

For example, a sample of typical entries for alternate names for Kabuli language will look like the following:

Code	Variant
00090	Afghan
00090	Farsi, Eastern: Dari: Kabuli
00090	Kabuli
00090	Pakhtoo
00090	Pakhtu
00090	Paktu
00090	Pashto
00090	Pashto: Kabuli
00090	Pashtu
00090	Pushto
00090	Pushtu
00090	Quetta-Kandahar Pashto

Note that some entries show both the language and the dialect, and include a colon and/or a comma. The second entry in the above table shows Farsi, Eastern: Dari: Kabuli. In this example, "Farsi, Eastern" is the equivalent of Eastern Farsi. The first colon separates the language name from the dialect, so "Dari" is the dialect. The second colon indicates that "Kabuli" is a sub-dialect of Dari.

Search by Name: The index table can be used to implement a search by name. For instance, the following query would return the five-digit codes for all the dialects that use the name *SomeName*

```
SELECT DISTINCT Code from ROD_AlternateNameIndex                    WHERE
Variant='SomeName'
```

Multiple dialects may be known by the same name. To verify the identity of a dialect, the user may consult the following link to the *Global Recordings Network* web site to see a report giving additional information about the selected dialect. For any ROD dialect code *ddddd* the following URL gives added information:

http://globalrecordings.net/dialect/ddddd

Additional information may be available on the Dialect Research Website as follows:
http://globalrecordings.net/research/dialect/ddddd

This web site contains a much larger selection of names, many of which have not yet qualified for inclusion in the ROD. For instance, the following link will show information for ASMAT: Waganu (dialect #04231.)

http://globalrecordings.net/research/dialect/04231

4. Change management

This section defines the process that the registry steward will follow to maintain the registry.

Governing philosophy

Rather than produce an exhaustive list of all possible dialects (varieties), the basic philosophy of the code set in this registry is to list only those dialects that have proven to be distinct enough to require different treatment in ministry strategies. Other varieties that still need to be investigated are listed in the Dialect Research Website.

Situations that will result in a change

There are a number of situations that could cause the Dialect Codes or their links to Language Codes to change. Situations include, but are not limited to the following:

(a) Changes that occur in the ISO 639-3 language code set may affect registered dialects.

(b) Speech varieties that are investigated and verified as ROD dialects according to the definition given in Section 1 will be assigned ROD Dialect Codes.

(c) Errors in the tables that are corrected according to the process outlined below may result in new codes being assigned.

How to make a change request

A change will be made when it can be shown to correct an error, or to improve the coverage and usefulness of the code set. If you believe any of the information in the Registry of Dialects is in error, or if you have an addition or improvement to suggest, send your proposed change by e-mail to RODsteward@globalrecordings.net Be sure to report the source of your information and include as much detail as possible.

How change requests are processed

The ROD steward will review the proposal and acknowledge receipt with initial comment or request for clarification. If the change appears to be in line with the philosophy of the registry, the steward will vet the proposal with others who are known to be using the registry. The data will also be added to the *Dialect Research Website*. It will be adopted only if there is a positive consensus among those consulted.

What to do pending a change

Each proposed dialect will be allotted a provisional five-digit code as soon as it is entered into the Dialect Research web site. Thus, when users need to use a code that is not yet part of a code set, they may freely use this provisional code until the outcome of a request to add a code is known.

How updates are made

The registry will be updated each February following the update of ISO 639-3 codes. Between official updates, the Dialect Research Web Site can be consulted for regular or provisional dialect codes.

5. Change history

-- See Next Page --

5. Change history

The registry will contain just one change history table.

<p align="center"><strong style="color:green">ROD_DialectChangeHistory</p>

All changes to the ROD_Dialect table are reported in ROD_DialectChangeHistory. This table is cumulative, listing all changes to successive versions of the registry. The table has the following four columns:

Code	char(5)	The dialect code that is affected by the change reported in this record.
Type	char(1)	A one-letter code indicating the type of change: There are four possible values: **C = Created**. The dialect code is newly created. **E = Extended**. The meaning of the code is extended by virtue of being merged with another code that has been retired. **M = Matching change**. Either the ROL code to which the Dialect was previously related has been changed, or additional research has indicated that the dialect is related to a different ROL code. **R = Retired**. The dialect code has been retired and should no longer be used in a database. **U = Updated**. There has been no change to the code or it's meaning, but other information in the code table entry (e.g. primary name, or the main country) has changed.
Date	char(10)	The date the change was released in a new version of the registry. Dates are expressed as 8 digits with hyphens to separate the parts of the date, e.g. YYYY-MM-DD.
Description	varchar(255)	Describes the change. In the case of **R** changes, it also describes what a user should do to fix existing data that uses the now retired code.

Note that there is not a change type for the case of narrowing the meaning of a code, such as when the dialect denoted by one code is split into two dialects. In such a case, the original code is retired, and two new codes are added. In this way, the user of the code set is assured that once a code has been used to tag an item of data, it will continue to be the right code to use for as long as the code remains an active member of the code set. Even after a code is retired, it will continue to have the same meaning as it had when it was last active, and though it is not recommended that a retired code continues to be used, it is not invalid to do so.

The SQL statement for creating the change history table is as follows:

```
CREATE TABLE ROD_DialectChangeHistory (
    Code          varchar(10) NOT NULL,
    Type          char(1) NOT NULL,
    Date          char(10) NOT NULL,
    Description   varchar(255) )
```

Here are some hypothetical examples or entries in the change history table:

Dialect Code	Type	Date	Description
10327	C	2005-12-07	Add Zemiaki
13980	R	2005-12-15	Same as 01432; change all 13980 to 01432
01432	E	2005-12-15	Includes 13980 that was retired
00681	R	2006-10-12	Merge with 14566; change all 00681 to 14566
00032	R	2007-08-09	Unable to verify existence; delete from database
11759	U	2008-01-10	Change name from Chikaranga to Karanga
12345	M	2008-05-11	ISO code split into two codes abd and ljg. 12345 now relates to abd

The change history table holds the cumulative list of all changes that have ever been made to the registry. Thus it may be queried to learn the complete history of a given code, or to learn all the changes that have been made since a given date. For instance, the following SQL query would be used to find out what changes have occurred since the beginning of 2005

```
SELECT * FROM ROD_DialectChangeHistory WHERE Date >=2005-01-01
```

For a site that has used ROL_Dialect codes in its own database, an important use of the change history table is to discover codes used in its data that are now obsolete and thus need to be changed. These will be only the codes that have been retired. Thus a full list of all data records needing to be changed can be found by doing a JOIN on the change history table. For instance, if the column named code in MyTable holds a ROD_Dialect code, then the following SQL statement will select all records that need to be changed due to changes to the code set since the beginning of 2005:

```
SELECT * FROM MyTable as M
JOIN ROD_DialectChangeHistory as C ON M.code=C.Code
WHERE C.Type='R' AND C.Date >=2005-01-01
```

Note that the Description field of the joined result set will describe what needs to be done to bring the selected language code up-to-date.