USING EIDR LANGUAGE CODES

Table of Contents

Introduction

Language codes appear in several places within the EIDR registry, generally either indicating the language expressed in a data field (for example, the language of a title string) or the language used in a piece of media (for example, the language spoken in a particular version of a work). To ensure universal comprehension and facilitate automated record de-duplication, EIDR uses a particular standard means of expressing language information.

EIDR language codes are quite flexible, and this can sometimes cause confusion. They allow for the designation of basic languages ("es" for Spanish) and regional dialects ("es-MX", the dialect of Spanish most often spoken in Mexico, as differentiated from "es-419", a form of Spanish that is intelligible throughout Latin America). It is also possible to designate the script used with a written language when a language might be expressed in different forms ("sr-Latn" for the Serbian language written using a Latin script as opposed to "sr-Cyrl" for Serbian written in Cyrillic). All of this is defined by BCP 47 (RFC 5646), *Tags for Identifying Languages*, published September 2009[1] as summarized in the LMT (Language Metadata Table), developed by MESA (Media & Entertainment Services Alliance) and standardized by SMPTE (Society of Motion Picture and Television Engineers).[2] This is the specification used by the XML lang tag, so EIDR language encodings can be validated using a standards-compliant XML parser. Such

[1] See https://tools.ietf.org/html/rfc5646.

[2] See https://www.mesalliance.org/language-metadata-table.

tags are case-insensitive, but to aid in human recognition, EIDR conforms to the casing conventions recommended in BCP 47.

Using BCP 47, it is possible to construct the same language code in different ways, though the shortest option is always preferred. Therefore, LMT specifies the preferred method of encoding each particular language, dialect, and script combination in common use within the media & entertainment industry. The EIDR Web UI provides the most common language codes as a drop-down selection list to aid data entry. If necessary, the user can select "Other" and manually enter a less common or more complex LMT language code.

NOTE: The one place EIDR language coding differs from LMT is in its treatment of the Chinese languages. In LMT, you code each Chinese dialect directly (yue, cmn, etc.), while EIDR prefixes them all with a generic Chinese language tage (zh-yue, zh-cmn). This way, a simple alphabetic sort groups together all variants of Chinese.

Recommended Data Entry Practice

Where possible, use the shortest valid language code for a given situation. If only the primary language family (or macro-language) is known for certain, do not guess at the possible extended language sub-tags or region codes. For example, if you know a work is in Chinese, but are not sure if it is Mandarin or Cantonese, do not guess. Code that as "zh".

NOTE: There is an implied precedence within the languages, so list the most important or common first, followed by any others in decreasing order of importance. The order has no impact on search or de-duplication; it only exists as a convenience for human review.

Original Language

Audio

Identify the most common or contextually important audible language(s) in the work as presented during the work's original release. If the region, dialect, or language extension is not a critical identifying characteristic of the work, then only include the primary language code.

NOTE: If it is important to further clarify the particular spoken language (as may be the case with Brazilian Portuguese or Quebecois), then add a suitable sub-tag to create a compound language tag ("pt-BR" for Brazilian Portuguese or "fr-CA" for Quebecois).

Visual

If visual languages are important to a work's narrative, such as when there is no significant audible language in the work, then code the most common or contextually important visual (generally, written) language(s) in the work as presented during the work's original release.

NOTE: In many cases, such works qualify as silent films and should be coded as described in *EIDR: Interim Best Practice – Silent Films*.

NOTE: In rare cases, the script used to represent a language may be an important distinguishing characteristic, for example "sr-Latn" vs. "sr-Cyrl" (Serbian written in Latin script vs. Serbian written in Cyrillic). These script codes are only valid for written languages and should be used only when necessary.

Version Language

A single work may have more than one language variant. In the EIDR system, these are recorded at the Edit and Manifestation levels. These language variants may be identified using Audio and/or Visual Version Language codes that differ from the Original Language codes. These are coded following the same practices as for Original Language.

Title, Alternate Title, Description

Identify the language used in the text field, irrespective of the audio or visual language(s) that may appear within the referenced work. For fanciful words that are not part of a specific language (such as *Jumanji*) or foreign words that have been borrowed into a language (such as *Ronin* into English), code the language based on the primary language of the work's original release (in both of these cases, "en" for English).

NOTE: If the characters used in the text field are not part of the standard script for that language, it may be helpful to identify the actual script used by appending a script tag to the language code. For example, use "ja-Hani" for Japanese expressed in Kanji characters. While it is valid to do so, it is not necessary to identify the script when a language has been transliterated into the Latin script (known as "Romanized"), since they can be identified by inspection. For example, Japanese transliterated into Latin script would just be "ja" rather than "ja-Latn". However, if it is important to identify the transliteration (for example, to distinguish otherwise similar subtitle tracks), then do so.

Constructing an EIDR Language Code

The general pattern for a BCP-47/LMT language code is:

language[-extended_language][-script][-region][-variant]

Where each sub-tag in square brackets is optional, but if present, separated from the preceding tag(s) by a hyphen. Each sub-tag may consist of a mix of letters or number, but no whitespace or punctuation (other than the separating hyphens) is allowed. BCP 47 allows a more complex language code structure, but EIDR best practice is to limit the codes to the format provided.

Possible language codes include:

- en – English
- fr-BE – The dialect of French common to Belgium
- fr-015 – The dialect of French commonly spoken across North Africa
- ja-Kana – Japanese written with Katakana script
- ja-Latn-hepburn – Japanese transliterated into the Latin alphabet using the Hepburn Romanization standard
- sgn-qmm – Mongolian Sign Language
- zh – Chinese
- zh-cmn – Mandarin Chinese

For more information on constructing language codes, and a list of all common language codes in Excel, PDF, or XML formats, please see https://www.mesalliance.org/language-metadata-table.