MDDF Language Tags
Encoding and Matching

This practice defines how to encode and match language codes.

REVISION HISTORY

Version	Date	Description
1.0	May 24, 2017	Initial publication

1 ENCODING LANGUAGE TAGS

Language shall be encoded in accordance with RFC 5646, *Tags for Identifying Languages* https://tools.ietf.org/html/rfc5646. The subtags that are available for use with RFC 5646 are available from the Internet Assigned Numbers Authority (IANA) at http://www.iana.org/assignments/language-subtag-registry.

The xs:language type shall be used for languages. Language should be as specific as possible; for example, 'ja-kata' is preferable to 'ja'.

Language tags can be confusing, but once you know the basic rules they are relatively straightforward.

A language tag is constructed using the following (from RFC 5646):

```
langtag       = language
                ["-" script]
                ["-" region]
                *("-" variant)
                *("-" extension)
                ["-" privateuse]
```

The details of each part are described in RFC 5646. As noted above, there must always be a `language` part; for example, 'fr' for French and 'en' for English. The `region` field is also commonly used, for example, 'fr-CA' (French, Canada) to represent Québécois. These other parts are called subtags. Language and subtag values can be found in the IANA Language Subtag Registry at http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry.

For example, within the registry, you could find:

```
Type: language
Subtag: fr
Description: French
Added: 2005-10-16
Suppress-Script: Latn
```

and

```
Type: region
Subtag: CA
Description: Canada
Added: 2005-10-16
```

These are the entries that respectively correspond with 'fr' and 'CA' in 'fr-CA'. Note that `Type` corresponds with the definition of langtag from RFC 5646.

Generally speaking, it is best not to be overly specific with language tags encoded in Avails or a Manifest unless complementary languages are provided. For example, if only one French language track is provided it is assumed that it will be used in France, Canada, Switzerland and other French speaking countries; so, it is best to encode it as 'fr' rather than 'fr-FR'. However, if multiple languages were provided, then be specific to differentiate them.

2 MATCHING LANGUAGE TAGS

When matching language tags, we are generally looking for the best fit to match a track or metadata to a user's language. For example, if the user's language is Québécois ('fr-CA') and the audio tracks are English ('en') and France French ('fr-FR'), what would be the best match?

The process for matching is well-defined and if everyone knows the rules, matching will be consistent and predictable.

The most obvious match is an identical match (e.g., 'fr-FR' to 'fr-FR'), but this isn't always possible. The trick is to pare down the language tags until the best match is found. As a rule, where there are multiple language tags using the same language subtag, use the best match ('fr' matches 'fr' better than 'fr-CA' and 'fr-CA' matches 'fr-CA' better than 'fr').

The process for matching is defined in RFC 4647, Matching of Language Tags, https://tools.ietf.org/html/rfc4647. The key is the have the right "Language Priority List" as defined in RFC4647, Section 2.3, https://tools.ietf.org/html/rfc4647#section-2.3. The Priority List is made for the user's language selection (e.g., browser language or system default language).

The assumed Priority List consists of at least the following language ranges:

1) The fully enumerated language tag including region, dialect or any other subtag element. For the 'fr-CA' user, this would be 'fr-CA'.

2) The language tag from the first entry trimmed to the primary language tag, followed by a wildcard '*' subtag. For the 'fr-CA' user, this would be 'fr-*'.

In this example, "fr-FR" Priority List will be "fr-FR, fr-*".

If the user's language is more specific, start with the full enumeration, then trim one subtag for each entry in the Priority list. For example, 'zh-Hant-CN' becomes 'zh-Hant-CN, zh-Hant, 'zh-*'.

The best language match between a language preference (e.g., System Language) and one or more languages in a list (e.g., language tags in a list of audio tracks) is to be done in accordance with [RFC4647], Section 3.4, "Lookup", https://tools.ietf.org/html/rfc4647#section-3.4. In short, work down the list until there is a match.

If no matches are found, it is suggested to use a default language. In Common Metadata, the default is indicated by the LocalizedInfo/@default attribute being 'true'.