Metadata Madness: *Will it ever STOP?*

Language Metadata Table

MESAlliance.org

Yonah Levenson (HBO), LMT Co-Chair

Wednesday, February 27, 2019
Agenda

- **Language Metadata Table Committee Introductions:**
  - Co Chairs: Yonah Levenson, Manager, & Laura Dawson, Metadata Analyst Metadata Management & Taxonomy @HBO
  - Working Group contributors include: Disney, Discovery, EIDR, European Union, HBO, Lionsgate, MESA, NBCUniversal, Paramount, Turner, Warner Bros, WWE, + vendors & many more

- Why LMT?
- Use Cases with LMT Solution
- LMT Working Committee Update
  - Mission Statement
  - Template for adding languages
- Next meeting: 3/13, 12:30-1:30 @HBO in NYC or concall
- Questions?
Language Metadata Table: Searching for the Lingua Franca

Common issues:

- Internationalization and localization are here; many depts have to define and track languages, including: Production, Marketing, Distribution, Legal, etc.
- Content often exists in more than one language
- Accessibility requirements abound
- System developers aren’t always familiar with metadata standards
  - Business asks for a new language value
  - Developers implement what was requested
  - Add to the mapping table(s)....
- LMT provides a unified standard of language terminology
Why IETF BCP-47?

- ISO 639 isn’t granular enough: Can’t handle Regional dialects
- ISO 639 is too granular: Can’t express broad geographic areas like Latin America
- The “Visual” or written language may be different from the Audio
  - Some languages expressed differently, inc. spellings. Ex: English, Chinese
  - Audio may have multiple dialects dependent upon the geographic region
- Language metadata codes are applied in many areas, including:
  - Audio
  - Visual or Written languages: Subtitles, Closed Captions, Audio description
  - User Interfaces
  - Rights and Licensing
  - Distribution
Solution: IETF BCP 47

- IETF: Internet Engineering Task Force
- BCP: Best Common Practice
- 47: The number of this best practice
- IETF BCP 47 consists of
  - ISO 639: Language codes
  - ISO 3166: Country codes
  - UN M. 49: UN Territory standards

- IETF BCP 47 works because
  - Language and geographic codes can be combined in more than 40K ways
  - Combine codes with territories for even more precision: “it-CH” = Italian as spoken in Switzerland
  - Updated language names reflect contemporary cultures: “Greenlandic” updated to “Kalaallisut”
  - A WWW standard supported by W3C
Use Cases and LMT Solution

“A language is a dialect with an army and a navy.”
- Max Weinreich, sociolinguist
Use Case 1: Spanish

- Spanish as spoken in Spain (Castilian) - es-ES
- Spanish as spoken in Mexico - es-MX
- Spanish as spoken in Latin America - es-419
Use Case 2: Chinese

- Cantonese Spoken Language: yue-CN
  - Audio
- Mandarin Spoken Language: cmn-CN
  - Audio
- Cantonese Written Language: zh-Hant
  - Subtitles CC UI
- Mandarin Written Language: zh-Hans
  - Subtitles CC UI
Use Case 3:
Italian/Neapolitan (My Brilliant Friend)
Language Metadata Table

Working Committee Update as of Feb 26, 2019
LMT Mission Statement (draft)

The Language Metadata Table standard was created to provide a unified source of reference for language codes for use throughout the broadcast and media industry. LMT’s mission is:

- To create a standardized table of language codes for implementation by entertainment and other industries using IETF BCP 47.
- To facilitate efficient and consistent LMT usage through best practices.
- To extend LMT code values through vetted field definitions and approved language code values with a community of thought leaders who focus on information and data from the business, professional associations and academic institutions through the exchange of knowledge and collaboration.
<table>
<thead>
<tr>
<th>Column Header Name</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language Grouping Name</td>
<td>The name of the language group, if appropriate. The Group name is equivalent to the generic language name. Language dialects are subordinate to their language grouping. Ex: Neapolitan falls under Italian.</td>
</tr>
<tr>
<td>Language Grouping Tag</td>
<td>IETF BCP 47 tag</td>
</tr>
<tr>
<td>Language Grouping Code</td>
<td>URN or URI for each language grouping value in the LMT.</td>
</tr>
<tr>
<td>Audio Language Tag</td>
<td>IETF BCP 47 language tag. Typically spoken/audio language.</td>
</tr>
<tr>
<td>Long Description 1</td>
<td>Description of language name in Latin script following IETF BCP 47 standard</td>
</tr>
<tr>
<td>Long Description 2</td>
<td>Alternate description of language name in Latin script following IETF BCP 47 standard</td>
</tr>
<tr>
<td>Audio Language Display Name 1</td>
<td>Endonym of written language. Typically the same as Audio Language Display Name 1 but not always.</td>
</tr>
<tr>
<td>Audio Language Display Name 2</td>
<td>Alternate endonym of written language. Typically the same as Audio Language Display Name 2 but not always.</td>
</tr>
<tr>
<td>Visual Language Tag 1</td>
<td>Script in which language is written following IETF-BCP-47 standard (which calls for the tags to be presented in Latin Script). Visual includes sign languages.</td>
</tr>
<tr>
<td>Visual Language Tag 2</td>
<td>Alternate script in which language is written following IETF-BCP-47 standard (which calls for the tags to be presented in Latin Script). Visual includes sign languages.</td>
</tr>
<tr>
<td>Visual Language Display Name 1</td>
<td>Endonym of written language. Typically the same as Audio Language Display Name 1 but not always.</td>
</tr>
<tr>
<td>Visual Language Display Name 2</td>
<td>Alternate written endonym. Typically the same as Audio Language Display Name 1 but not always.</td>
</tr>
<tr>
<td>Code</td>
<td>URN or URI for each language value in the LMT.</td>
</tr>
<tr>
<td>Column Header Name</td>
<td>Example 1: Serbian</td>
</tr>
<tr>
<td>----------------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>Language Grouping Name</td>
<td>Serbo-Croatian</td>
</tr>
<tr>
<td>Language Grouping Tag</td>
<td>sh</td>
</tr>
<tr>
<td>Audio Language Tag</td>
<td>sr</td>
</tr>
<tr>
<td>Long Description 1</td>
<td>Serbian</td>
</tr>
<tr>
<td>Long Description 2</td>
<td></td>
</tr>
<tr>
<td>Audio Language Display Name 1</td>
<td>Srpski</td>
</tr>
<tr>
<td>Audio Language Display Name 2</td>
<td>српска</td>
</tr>
<tr>
<td>Visual Language Tag 1</td>
<td>sr-Latn-RS</td>
</tr>
<tr>
<td>Visual Language Tag 2</td>
<td>sr-Cyrl-RS</td>
</tr>
<tr>
<td>Visual Language Display Name 1</td>
<td>Srpski</td>
</tr>
<tr>
<td>Visual Language Display Name 2</td>
<td>српска</td>
</tr>
<tr>
<td></td>
<td>urn:ietf:bp47:sh-RS</td>
</tr>
</tbody>
</table>
LMT Working Committee Agenda:
3:30 Today!!

- Mission Statement draft review
- Column Head Definitions: Change requests
  - Visual to *Written or Signed*
    - Note: Gallaudet has approved Visual as it covers Sign Language
  - Audio to *Verbal*
  - Shorten Language to *Lang*
- Audio Language Display Name 1 definition change:
  - *Endonym of written language. Typically the same as Visual Language Display Name 1 but not always.*
- Audio Language Display Name 2 definition change:
  - *Endonym of written language. Typically the same as Visual Language Display Name 2 but not always.*
Working Committee Agenda (cont): Today!!

- Language Grouping Tag to *Language Top Grouping*
- Code to *Language Code*
- Additional language requests
  - 50 from Disney
  - Using draft template
  - Sign language: which languages to include for starters?
- Policies and Procedures
  - Submission process
  - Formats
- March meeting:
  - @HBO in NYC, March 13 12:30-1:30
- Next steps
Language Groupings: Think about

• What do you do when you know it’s language X, but not which flavor of X?
  • Dialect difference?
  • When the dialect has a navy, so it’s officially a language difference?
• Common examples: Chinese, Spanish, Portuguese, French, Sign Language
• EIDR’s proposal for alternate language family encoding: i.e., “zh-yue” instead of “yue”
• Identifying language families in the LMT spreadsheet
LMT Language Grouping Proposal

- Use IETF BPC 47 "Macrolanguage" and "Language Family" designations
- Allows for alphabetical sort by grouping, keeping languages like Chinese together
  - otherwise, Mandarin and Cantonese would separate
- Simple hierarchy allows for maximum flexibility
Language Grouping Examples

- **Greek**: to account for ancient vs modern
- **English**: British, Canadian, Australian, American, etc.
- **Spanish**: Latin American vs European
- **Chinese**: Mandarin vs Cantonese vs Min Nan, etc.
- **Sign Languages**
- **Special**: for “undetermined” and “no linguistic content”
Summary

- IETF BCP 47 provides the most flexibility for capturing language metadata because it’s a Standard of Standards
  - Extensible
  - Capture what is needed for your business need
  - Document the solution
  - Implement across the Enterprise
  - Encourage others in the industry to adopt IETF BCP 47 by sharing the approach
  - Update values as needed

- LMT working committee is moving forward
  - Meeting today @3:30
  - Meeting in NYC at HBO on 3/13 at 12:30, or online
  - Goal is to be in maintenance mode for adding languages going forward
  - HBO is maintaining LMT in its taxonomy tool; output available via MESA