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

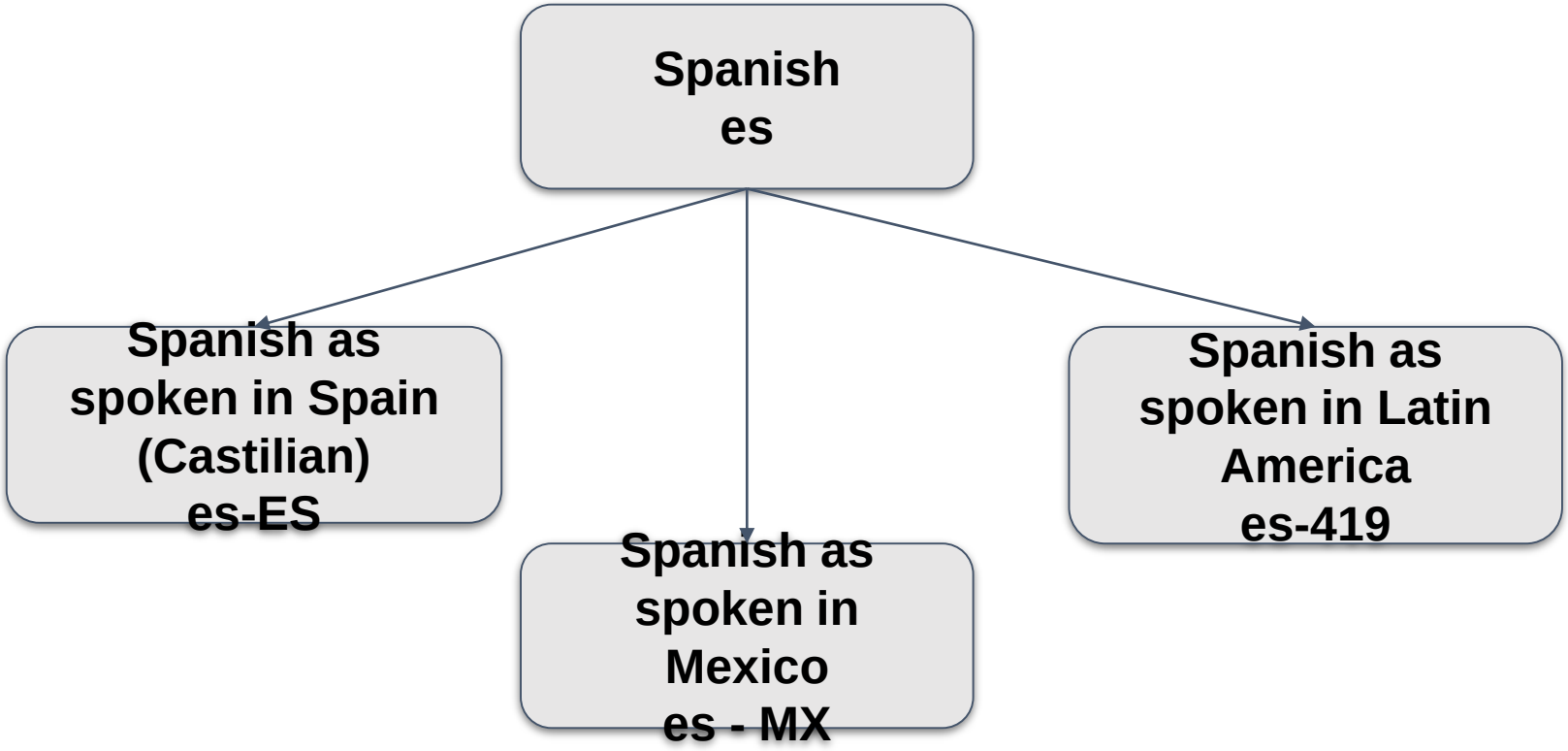
Solution: IETF BCP 47

- 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

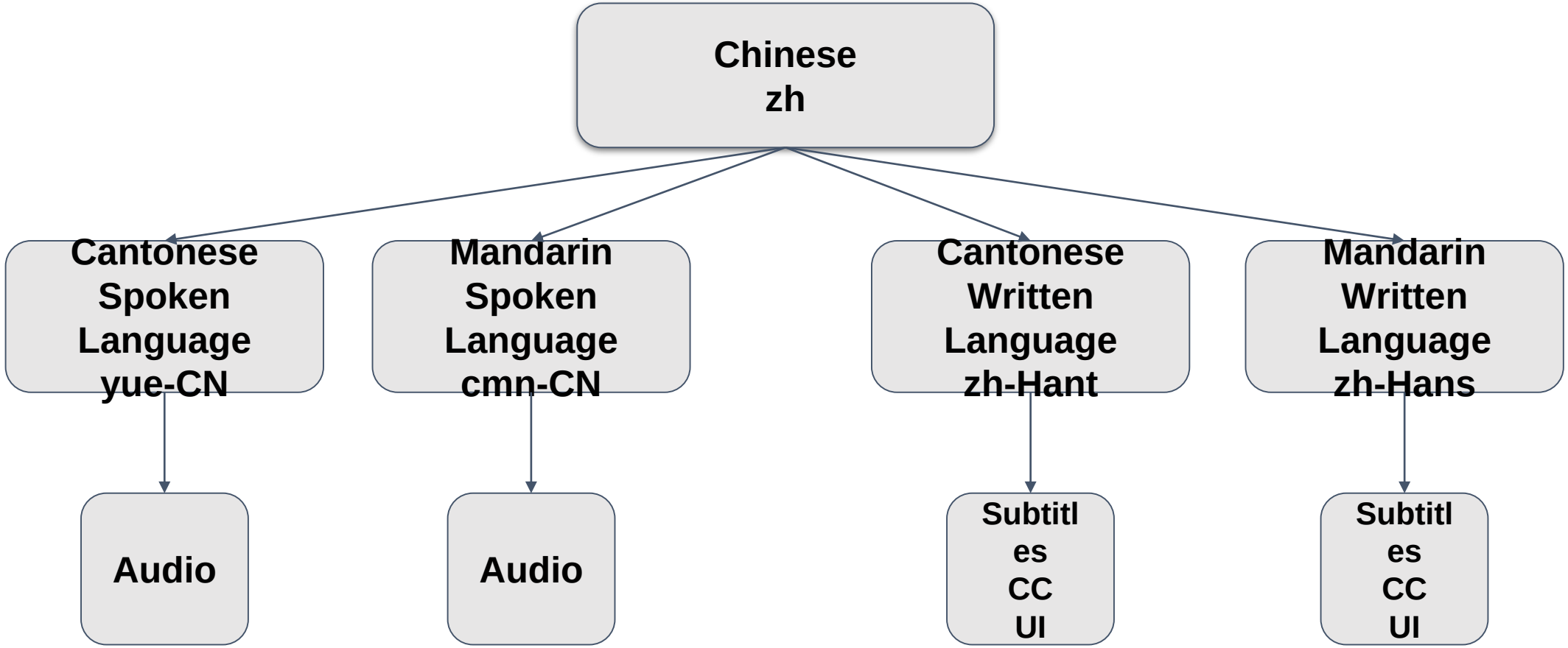
Use Cases and LMT Solution

“A language is a dialect with an army and a navy.”
- Max Weinreich, sociolinguist

Use Case 1: Spanish



Use Case 2: Chinese



Use Case 3:
Italian/Neapolitan (My Brilliant Friend)

Italian
it

Neapolitan
nap

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.

Template: Additional Languages

Column Header Name	Definition
Language Grouping Name	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: Neopolitan falls under Italian.
Language Grouping Tag	IETF BCP 47 tag
Language Grouping Code	URN or URI for each language grouping value in the LMT.
Audio Language Tag	IETF BCP 47 language tag. Typically spoken/audio language.
Long Description 1	Description of language name in Latin script following IETF BCP 47 standard
Long Description 2	Alternate description of language name in Latin script following IETF BCP 47 standard
Audio Language Display Name 1	Endonym of written language. Typically the same as Audio Language Display Name 1 but not always.
Audio Language Display Name 2	Alternate endonym of written language. Typically the same as Audio Language Display Name 2 but not always.
Visual Language Tag 1	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.
Visual Language Tag 2	Alternate script in which language is written following IETF-BCP-47 standard (which calls for the tags to presented in Latin Script). Visual includes sign languages.
Visual Language Display Name 1	Endonym of written language. Typically the same as Audio Language Display Name 1 but not always.
Visual Language Display Name 2	Alternate written endonym. Typically the same as Audio Language Display Name 1 but not always.
Code	URN or URI for each language value in the LMT.

Template: Populated Examples

Column Header Name	Example1: Serbian	Example 2: Mandarin (spoken)	Example 3: Armenian - Eastern	Example 4: Armenian - Western
Language Grouping Name	Serbo-Croatian	Chinese	Armenian Family	Armenian Family
Language Grouping Tag	sh	zh	hyx	hyx
Language Grouping Code	urn:ietf:bcp:47:sh	urn:ietf:bcp:47:zh	urn:ietf:bcp:47:hyx	urn:ietf:bcp:47:hyx
Audio Language Tag	sr	cmn	hy	hyw
Long Description 1	Serbian	Mandarin	Armenian	Armenian as spoken by the Armenian Diaspora
Long Description 2				
Audio Language Display Name 1	Srpski	普通话	հայերէն	հայերեն
Audio Language Display Name 2	српска			
Visual Language Tag 1	sr-Latn-RS	zh-Hans	hy	hyw
Visual Language Tag 2	sr-Cyrl-RS			
Visual Language Display Name 1	Srpski	简体中文	հայերէն	հայերեն
Visual Language Display Name 2	српска			
Code	urn:ietf:bcp:47:sr	urn:ietf:bcp:47:cmn	urn:ietf:bcp:47:hy	urn:ietf:bcp:47:hy
	urn:ietf:bcp:47:sr-Latn-RS	urn:ietf:bcp:47:cmn-CN	urn:ietf:bcp:47:hy-AM	urn:ietf:bcp:47:hy-US
	urn:ietf:bcp:47:sr-Cyrl-RS	urn:ietf:bcp:47:zh-Hans		
	urn:ietf:bcp:47:sr-RS			

MESA

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