MESA
Language Metadata Table (LMT) Policies and Best Practices

Documentation 1.2

Version History

Date	Description	Creator/Editor	Version
2022-08-07	Structural edits ; Updated URL codes in Language Groups ; Updated examples	Jaclyn Zepnick	1.2
2019-10-04	Added “non-audio languages” section	Laura Dawson	1.1
2019-07-22	Minor tweaks	Yonah Levenson, Laura Dawson	1.0
6/25/2019	Added language from Craig Seidel re: compliance and governance	Yonah Levenson	0.3
6/19/2019	Incorporated updates/ feedback and fixed formatting	Yonah Levenson	0.2
5/14/2019	Initial Draft	Yonah Levenson, Laura Dawson, Colleen Henderson, HBO	0.1

Table of Contents

1. Introduction

1.1. What is MESA?

MESA, or The Media & Entertainment Services Alliance, is "focused on three core M&E technologies: data, IT and security. MESA's 120-plus members and content advisors collaborate to advance change management, new workflow solutions and production/supply chain efficiencies".

MESA's five technology communities include:

- The Hollywood IT Society
- The Smart Content Council
- Content Delivery & Service Association
- Smart Screen
- Women in Technology: Hollywood

1.2. What is the LMT?

The Language Metadata Table (LMT) is an expandable mapping resource that is used to organize language metadata via locations and dialects. It was created to provide a unified source of reference for language codes for use throughout the media and entertainment industries. The group of people involved in developing and incorporating the LMT across companies allows for industry-wide collaboration. The goal is to unite data specialists under a single "open source" table of language metadata values for the media and entertainment industry.

The LMT was conceived in 2017 at WarnerMedia's HBO by its Metadata Management and Taxonomy department. Extensive analysis was conducted on HBO's existing language values in multiple systems, as well as on industry standards. The conclusion was that HBO should use the IETF BCP 47 standard – defined below:

- Internet Engineering Task Force (IETF), a voluntary trade association responsible for TCP/IP standards and other internet infrastructure
- Best Current Practice (BCP)
- 47: IETF BCP's 47th best current practice

LMT adheres to IETF BCP 47.

MESA supports the LMT Working Group as one of its communities of practice.

HBO was asked by MESA to share their internal LMT with the media and

entertainment industry which they did in July 2018 at MESA's Smart Content Summit in NYC.

HBO encouraged industry adoption of the LMT/IETF BCP 47 languages standard and, in the spirit of industry collaboration, granted MESA the rights to publish HBO's work and continue the efforts of the LMT Working Group.

The LMT Working Group was established and since its initial release and includes members from the following organizations: HBO, Amazon Studios, Warner Bros, The Walt Disney Company, NBCUniversal, Paramount, Sony Pictures, Lionsgate, Showtime, among others. These representatives have reached agreement on:

- Language terms, definitions, and examples
- Template for adding new languages
- Over 200 language codes and display values – with ongoing analysis and research to include more as use cases arise

In addition to studios, vendors, Subject Matter Experts (SMEs), and standards organizations such as EIDR, participate in the ongoing expansion of the LMT.

1.3. The LMT Mission Statement

The LMT standard was created to provide a unified source of reference for language codes to use throughout the Media and Entertainment industries. 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

2. LMT Background

2.1. The Creation of the LMT

The LMT initiative began at HBO in 2017. Like many enterprises, HBO has multiple systems that are used across the enterprise, including: production systems, marketing systems, scheduling systems, etc. The Metadata and Taxonomy team was

asked what the language codes for Spanish as spoken in Latin America should be. The team's research showed that across these arious internal systems, not one system used the same code – "SPA-LA", "LAS", "LATAM SPA", etc. Further analysis discovered that each language had multiple language codes.

A project was created to normalize language codes across the enterprise. Thus, the LMT was created. The initial table had 128 languages.

2.2. Approach

IETF (Internet Engineering Task Force) language codes are recommended by the World Wide Web Consortium for encoding languages in HTML, CSS, XML, JSON and other data transmission formats and markup languages. The language codes are referred to as IETF BCP 47 (Best Current Practice 47).

IETF BCP 47 incorporates numerous ISO language and territory standards in a precise combination. For those regions that have no ISO 3166 equivalent, IETF BCP 47 utilizes the UN M49 territory standard.

When creating HBO's internal language standard, the Metadata and Taxonomy team investigated a number of international language standards, including ISO 639 1, 639 2, 639 3, and IETF BCP 47, which resulted in the following findings:

- ISO 639 is not granular enough – it is unable to handle regional dialects
- Yet, ISO 639 can be too granular – unable to express broad geographic areas like Latin America
- Current language standards do not account for the differences between visual and written languages – the code for the written language may differ from its spoken counterpart. Additionally, there may be multiple spoken dialects corresponding with only one or two written languages.
    - For example, in Chinese there are only two forms of written scripts vs. various spoken dialects
    - Audio languages may have multiple dialects depending on the geographic region

The final decision was to implement IETF BCP 47 language codes to achieve the required granularity of languages codes needed at HBO.

IETF BCP 47 is based on the following existing metadata standards (note that the LMT has only used values from the first three bulleted standards):

- ISO 639: two and three-character language codes

- ISO 3166: two-character country codes
- UN M.49: United Nations three-digit numeric territory codes
- ISO 15924: four-character script codes

The complete code syntax looks like the following:

- Language-script-region-variant-extension-privateuse

For example, “mn-Cyrl-MN” represents Mongolian written in Cyrillic as used in Mongolia.

IETF BCP 47 works for the following reasons:

- There are 40K+ language, script, and geographic codes which can be combined in an exponential number of ways
- It is possible to combine language codes with territories to allow for even more precision, ex: “it-CH” = Italian as spoken in Switzerland
- Updated language names reflect contemporary cultures (i.e. “Greenlandic” updated to “Kalaallisut”)
- It is a WWW standard supported by W3C

As stated, IETF BCP 47 codes can be combined for greater descriptive granularity. For example, the code for “English as spoken in the US” is “en-US”. Whereas the code for “English as spoken in Great Britain” is “en-GB”. This is helpful in accessibility applications where it is key for the hearing-impaired user to understand the context of a characters’ speech.

Unlike ISO 639, IETF BCP 47 was not developed as a bibliographic standard, though it has bibliographic applications. IETF BCP 47 was developed to describe languages within internet applications on the web. Thus, it is more suited to the purpose of describing the languages of digital assets than the ISO standard.

Additionally, IETF BCP 47 provides for description of fictional languages (ex: Klingon). Meaning it can accommodate languages such as HBO’s Dothraki from *Game of Thrones*, as well as other invented languages. IETF BCP 47 is a proven solution to future-proof the language metadata requirements that are needed within the Media and Entertainment industry.

3. Common Language Metadata Needs Across the Industry

As content becomes increasingly global, the Media and Entertainment industry requires standardized language codes for the following:

- Audio

- Visual or written languages
    - Subtitles
    - Closed Captions
    - Burned In Captions/Forced Narrative
    - User Interfaces
- Rights and Licensing
- Distribution
- Accessibility
    - Audio description/descriptive narration for the visually impaired
    - Sign language interpretation

The LMT includes codes that can be applied for each language need.

4. LMT Language Grouping

Language groupings are an optional and useful way to work with LMT. The use of IETF BCP 47 "Macrolanguage" and "Language Family" designations allow for alphabetical sorting by grouping, keeping languages like Chinese together. If not, languages like Mandarin and Cantonese would be separate. A simple hierarchy allows for the maximum flexibility. The following are language grouping examples:

- Greek – to account for Ancient vs. Modern
- English – differentiations between British, Canadian, Australian, American, etc.
- Spanish – differentiations between Latin American vs European, Mexican vs Argentinian, etc.

List of Language Groups and their metadata:

Language Group Name	**Language Group Tag**	**Language Group URL**
Akan	ak	https://smpte-ra.org/register/lmt/group/ak
Albanian	sql	https://smpte-ra.org/register/lmt/group/sqj
Algic	aql	https://smpte-ra.org/register/lmt/group/aql
Arabic	ar	https://smpte-ra.org/register/lmt/group/ar
Armenian	hyx	https://smpte-ra.org/register/lmt/

		group/hyx
Bantu	bnt	https://smpte-ra.org/register/lmt/group/bnt
Basque	euq	https://smpte-ra.org/register/lmt/group/euq
Catalan	ca	https://smpte-ra.org/register/lmt/group/ca
Chinese	zh	https://smpte-ra.org/register/lmt/group/zh
Dutch	nl	https://smpte-ra.org/register/lmt/group/nl
Estonian	et	https://smpte-ra.org/register/lmt/group/et
French	fr	https://smpte-ra.org/register/lmt/group/fr
German	de	https://smpte-ra.org/register/lmt/group/de
Greek	el	https://smpte-ra.org/register/lmt/group/el
Gujarati	gu	https://smpte-ra.org/register/lmt/group/gu
Hindi	hi	https://smpte-ra.org/register/lmt/group/hi
Italian	it	https://smpte-ra.org/register/lmt/group/it
Japanese	ja	https://smpte-ra.org/register/lmt/group/ja
Kannada	kn	https://smpte-ra.org/register/lmt/group/kn
Latvian	lv	https://smpte-ra.org/register/lmt/group/lv

Malagasy	mg	https://smpte-ra.org/register/lmt/group/mg
Malay	ms	https://smpte-ra.org/register/lmt/group/ms
Malayalam	ml	https://smpte-ra.org/register/lmt/group/ml
Marathi	mr	https://smpte-ra.org/register/lmt/group/mr
Mon-Khmer	mkh	https://smpte-ra.org/register/lmt/group/mkh
Mongolian	mn	https://smpte-ra.org/register/lmt/group/mn
Norwegian	no	https://smpte-ra.org/register/lmt/group/no
Odia	or	https://smpte-ra.org/register/lmt/group/or
Odia	or	https://smpte-ra.org/register/lmt/group/or
Pashto	ps	https://smpte-ra.org/register/lmt/group/ps
Persian	fa	https://smpte-ra.org/register/lmt/group/fa
Portuguese	pt	https://smpte-ra.org/register/lmt/group/pt
Serbo-Croatian	sh	https://smpte-ra.org/register/lmt/group/sh
Sign Languages	sgn	https://smpte-ra.org/register/lmt/group/sgn
Spanish	es	https://smpte-ra.org/register/lmt/group/es

Swahili	sw	https://smpte-ra.org/register/lmt/group/sw
Tamil	ta	https://smpte-ra.org/register/lmt/group/ta
Thai	th	https://smpte-ra.org/register/lmt/group/th
Uto-Aztecan	azc	https://smpte-ra.org/register/lmt/code/azc
Uzbek	uz	https://smpte-ra.org/register/lmt/group/uz

Figure 1: LMT Language Grouping Table

5. IETF BCP 47 Rules

According to the BCP 47 standard, "A language tag is composed from a sequence of one or more 'subtags', each of which refines or narrows the range of language identified by the overall tag."

This sequence of subtags must be created in the following order:

- **Language** – a short code (two or three letters) in lowercase
- **Script** – first letter of the tag is capitalized, with lowercase ensuing letters
- **Region** – all capital letters, unless the region code is from UN M49, which consists of numbers
- **Variant** – indicates an orthographic, historical, or defined dialect version of the primary language, and can be alphanumeric
- **Extension** – preceded by a single lowercase letter, and used to generate identifiers for languages; extension subtags consist of two to eight lowercase letters
- **Private Use** – typically preceded by an "x", these tags are for use in situations specific to private agreements, and are agreed on by both parties to the agreement (i.e. for internal-only codes)

Up until July 2022, only the first three subtags (Language, Script, Region) were in use. However, the Working Group approved that LMT is to provide recommendations for private-use codes as use cases arise.

Examples of combinations:

Code	Language Description

zh-Hans	Chinese written in Simplified Script
vls	Flemish, (AKA Dutch as spoken in Belgium)
ja-Jpan-JP	Japanese written with Han, Hiragana, and Katakana characters
sr-Latn	Serbian written in Latin script

Figure 2: LMT Language Combination Examples

IETF BCP 47 specifies that the shortest possible tag should be used.

If a language is being described without the context of the country in which it is spoken, only the primary (first) tag should be used. A use case for primary-only tags is in rights negotiations where a content company negotiating for the rights to distribute Spanish content is not necessarily going to distinguish among the various types of Spanish that exist. In this instance, the simple code “es” would be used in systems that record this process.

However, on the distribution side, it is crucial to inform trading partners and consumers which variation of Spanish they will be listening to or reading. Thus, additional qualifiers would be necessary in those systems.

6. LMT Template

This section contains information about the LMT template, including definitions and best practices for implementation and submission. The LMT template is available for download on the LMT website.

6.1. Template Definitions

Column Header Name	Definition
Language Group 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: Armenian - Western falls under the Armenian group.
Language Group Tag	IETF BCP 47 tag
Language Group URL	URL for each language group value in the LMT. This URL will* be used to validate the language tags in the SMPTE validator.

Audio Language Tag	IETF BCP 47 language tag for spoken/audio language
Long Description 1	Description of language name in Latin script that follows the IETF BCP 47 standard
Long Description 2	Alternate description of language name in Latin script that follows the IETF BCP 47 standard
Audio Language Display Name 1	Endonym of audio language. Typically the same as Visual Language Display Name 1, but not always
Audio Language Display Name 2	Alternate endonym of audio language. Typically the same as Visual Language Display Name 2 but not always
Visual Language Tag 1	Script in which language is written that follows the IETF BCP 47 standard, which calls for tags to be presented in Latin script
Visual Language Tag 2	Alternate script in which language is written that follows the IETF BCP 47 standard, which calls for tags to be presented in Latin script
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 2, but not always
URL	URL for each language value in the LMT. This URL will* be used to validate the language tags in the SMPTE validator.

Figure 3: LMT Template Terms and Definitions

*Pending SMPTE approval

6.2. Populated Examples

Column Header Name	**Example 1: English**	**Example 2: Spanish**	**Example 3: Serbian**	**Example 4: Mandarin**
Language Group Name	English	Spanish	Serbo-Croatian	Chinese
Language Group Tag	en	es	sh	zh
Language Group URL	https://smpte-ra.org/register/lmt/group/en	https://smpte-ra.org/register/lmt/group/es	https://smpte-ra.org/register/lmt/group/sh	https://smpte-ra.org/register/lmt/group/zh
Audio Language Tag	en	es-419	sh	cmn
Long Description 1	English	Spanish as spoken in Latin America	Serbo-Croatian	Mandarin
Long Description 2				
Audio Language Display Name 1	English	Español	srpskohrvatski	普通话
Audio Language Display Name 2			српскохрватски	国语
Visual Language	en	es-419	sh-Latn	

Tag 1				
Visual Language Tag 2			sh-Cyrl	
Visual Language Display Name 1	English CC	Español	srpskohrvatski	
Visual Language Display Name 2			српскохрватски	
URL	https://smpte-ra.org/register/lmt/code/en	https://smpte-ra.org/register/lmt/code/es-419	https://smpte-ra.org/register/lmt/code/sh	https://smpte-ra.org/register/lmt/code/zh-cmn

Figure 4: LMT Examples: Populated as Template Entries

7. Language Update and/or Addition Processes

Any party can submit use cases to MESA's LMT Working Group. These submissions will be reviewed by the LMT Co-Chairs and presented to the Working Group to decide whether the requests are incorporated into the LMT.

There is always a need for updates to existing languages or requests for new languages as languages are in continual flux for various reasons including political factors.

When submitting an update or requesting a new language to be added, the following steps are required:

- Download the submission template from the LMT website
- Populate the template following IETF BCP 47 rules (https://www.rfc-editor.org/info/bcp47#section-1)
    - Research may be required and the LMT Co-Chairs ask that the requester provide that research at time of submission
- Submit to the LMT Co-Chairs for initial review (lmt@mesaonline.org)
- The request(s) will be shared with the LMT Working Group for review, feedback, and formal approval
- Upon approval, the changes and/or additions will be added to the LMT and

the updated table will be available to download on the website
- With each new release, LMT members will receive an email describing the changes and/or additions.

8. Media & Entertainment Use Cases

This section contains definitions and examples of where and why languages need to be captured within the Media and Entertainment industry.

8.1. Audio

There is an industry-wide need to describe languages used in audio tracks for use in communicating what the audio track language consists of.

Use Case – to ensure that the correct language audio track corresponds to the requirements of affiliates broadcasting in that language.

Ex: “es-419” (Spanish as spoken in Latin America)

8.2. Closed Captions

There is an industry-wide need to be able to offer closed captioning, where the end user can turn them off and on. There is a need to describe closed captions with language metadata such as language type and script/writing system.

Use Case – to remain compliant with the Americans with Disabilities Act (ADA) and Federal Communications Commission (FCC) requirements.

Ex: “sr-Cyrl” (Serbian as written in Cyrillic)

8.3. “Burned In” Values/Subtitles

For text that is not user-controlled, but rather “burned in” to the video, there is an industry wide need to provide descriptions using language metadata such as language type and script/writing system.

Use Case – to ensure affiliates are broadcasting the correct content, and that users see the writing systems they expect.

Ex: “zh-Hant-HK” (Hong Kong Chinese in Traditional Script)

8.4. Accessibility

There is an industry-wide need to be able to offer “visual description”, or a

narration of what occurs on the screen for the visually-impaired.

Use Case – to remain compliant with the Americans with Disabilities Act (ADA) and Federal Communications Commission (FCC)requirements.

Ex: “pt-BR” (Portuguese as spoken in Brazil)

8.5. Acquisition and Rights

When acquiring rights to new content, organizations may not necessarily be concerned with the granularity of language data. For example, rights are acquired to broadcast in the overarching language of Spanish, versus a particular variation of Spanish.

Use Case – to describe an organization’s right to distribute content in a specific language.

Ex: es (Spanish)

8.6. Trading Partners and Consumer Viewing

Affiliates, trading partners, and consumers need to know the full description of the content they are receiving, including a granular description of the language as it is spoken within the content. Thus, the geo-specific tags are implemented in these instances.

Use Case – to convey what specific language the content is in.

Ex: es-ES (Castilian, AKA Spanish as spoken in Spain)

8.7. Non-Audio Languages

There are several languages which are not spoken and thus do not have an audio component. For LMT’s purposes, these include sign languages, as well as Norwegian Bokmål, the Norwegian written standard adopted by 85-90% of the country.

Language Grouping	Language Grouping Code	Audio Language Tag	Written Language Tag	Language Name
Sign Language	sgn		ase	American Sign Language

Sign Language	sgn		asf	Australian Sign Language
Sign Language	sgn		bfi	British Sign Language
Norwegian	no		nb	Norwegian Bokmål

Figure 5: Languages without Audio

9. Examples and Use Cases

This section contains specific examples and use cases within different elements of the Media and Entertainment industry.

Example 1: Spanish

The code in the top box, “es”, could be used for acquisition and rights purposes. The codes in the bottom boxes could be used for distribution to trading partners, affiliates, and consumers.

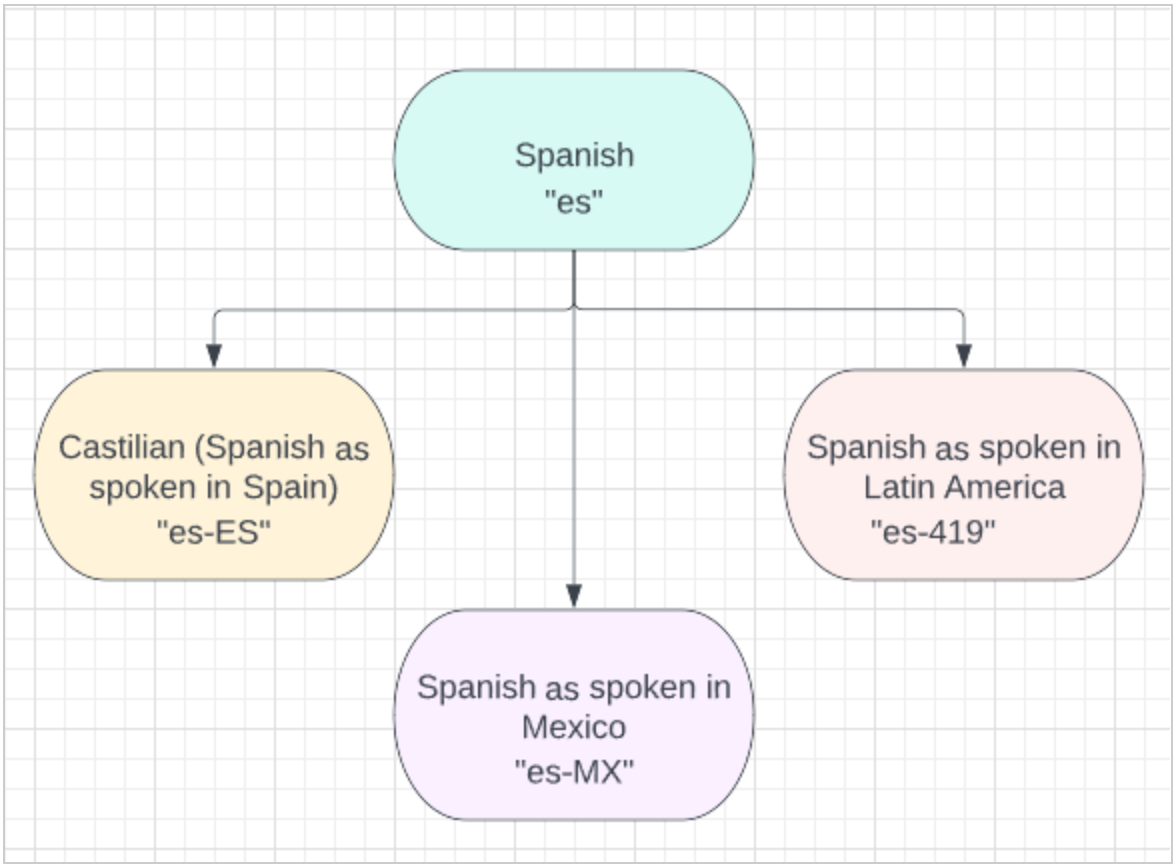

Figure 6: Spanish LMT Example: Generic, Latin America, Spain, and Mexico

Example 2: Chinese

The code in the top box, “zh”, could be used for rights and acquisitions purposes. The codes in the bottom left boxes could be used to describe audio content to trading partners, affiliates, and consumers. The codes in the bottom right boxes could be used to describe written content (subtitles, etc.) to trading partners, affiliates, and consumers.

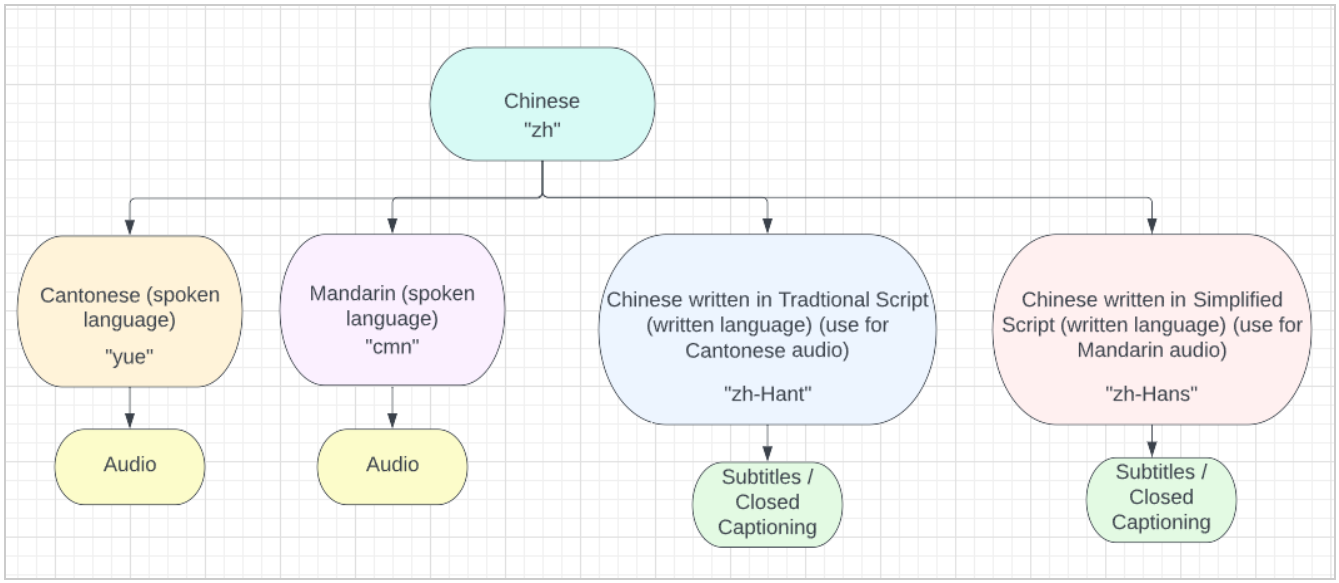

Figure 7: Chinese example for spoken and written languages.

Example 3: Italian / Neapolitan

The use case for this example is for the series, *My Brilliant Friend*.

The following is an example of a language, Neapolitan, that has historically been considered a dialect of Italian. Aside from a limited number of “variant” subtags, IETF BCP 47 makes no distinctions between dialects and languages. Thus, the Italian macrolanguage tag is not included in the Neapolitan language tag.

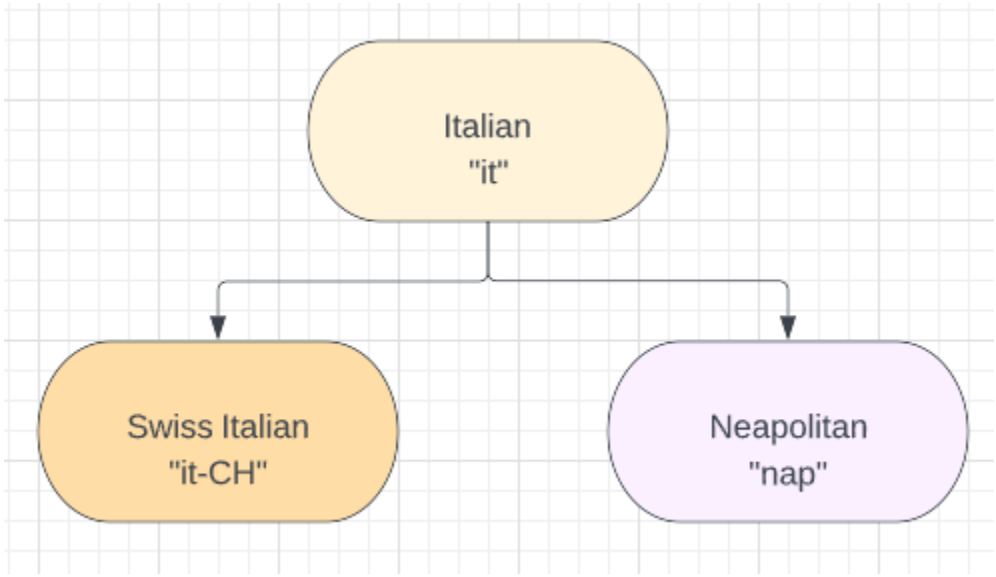

Figure 8: Italian and Neapolitan Language Code Examples.

10. Compliance Best Practices

Compliance with LMT requires the use of LMT tags and/or URLs for languages covered by LMT use cases and that are included in the current, published LMT. That is, where LMT has defined one or more language tags and URLs for a use case in the published table, those tags and/or URLs must be used.

The LMT Working Group recognized that use cases not covered by LMT might require the use of codes not in LMT.

- When a use case has not been considered by LMT, it is acceptable to use tags or URLs that are not in the list
- It is expected that those use cases will be submitted to LMT as soon as practical
- If the LMT Working Group has rejected a use case, those tags and URLs should not be used

It is possible that existing and/or legacy systems may not be able to comply with LMT codes due to a variety of reasons. If those systems cannot comply, the onus is on the system owner to create and maintain their own mapping table and/or other tools to transform the LMT codes into a code that their system(s) can handle.

Additionally, if exports of language codes are required from these systems, the export shall deliver LMT compliant codes. The maintenance of the mapping table(s) and tools is outside of the scope of the LMT Working Group.

11. Conclusions

LMT's solution for language is to implement IETF BCP 47. These codes are future-proof, and meet the wide variety of language requirements that the Media and Entertainment industry experiences.

The BCP 47 language codes from IETF are used in a wide variety of applications. They are flexible, granular, and modular, such that they can describe language contexts both broad and narrow. They are based on ISO and UN codes and are highly stable and widely adopted by other standards bodies such as W3C.

Additionally, IETF BCP 47 provides for distinctions between spoken and written language via the "Script" subtag. This allows for the ability to code close-captioning and subtitling in addition to spoken dialogue.

The decision to follow IETF BCP 47 standards is based on the conclusion that it has the most flexibility for capturing language metadata. Moving forward, the LMT Working Group will continue to meet on a regular basis to discuss policy and procedures, ultimately collaborating to expand the LMT based on relevant use cases that allow for efficient content distribution across the Media and Entertainment industry and beyond.

12. Appendix

12.1. Resources

- BCP 47
- IETF BCP 47 Complete Language Registry
- ISO 639
- ISO 639-2 Library of Congress
- UN M49

12.2. Media

- HBO Looks to Demystify Language Metadata
- In Collaboration With MESA, SMPTE Brings First-of-a-Kind Language Metadata Table Register Into Its Agile New Public CD Process

12.3. Contact Information

LMT Co-Chairs:

- Yonah Levonson
- Meg Morrissey
- Jaclyn Zepnick

  Co-Chair email: lmt@mesaonline.org