MESAlliance

Language Metadata Table (LMT) Policies and Best Practices Documentation 1.1

Documentation Updates

Date	Description	Creator/Editor	Version
2019-10-04	Added “non-audio languages” section	Laura Dawson	1.1
2019-07-22	Minor tweaks	Yonah Levenson, Laura Dawson	1.0 released
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 MESAlliance?

The MESAlliance, or The Media & Entertainment Services Alliance, is "focused on three core M&E technologies: data, IT and security. MESAlliance's 120-plus members and content advisors collaborate to advance change management, new workflow solutions and production/supply chain efficiencies".[1] Their five technology communities include: the Hollywood IT Society, the Smart Content Council, Content Delivery & Service Association, Smart Screen, and Women in Technology: Hollywood.

1.2 What is the LMT?

The Language Metadata Table[2] (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 expanding the LMT across companies allows for collaboration on what can be used industry wide, to unite data specialists under a single "open-source" table of language metadata values for the media and entertainment industry.

The LMT was conceived at 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 IETF BCP 47.

- 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.

HBO was asked by MESAlliance to share their internal LMT with the media and entertainment industry which they did on July, 2018 at MESAlliance's Smart Content Summit in NYC. MESAlliance supports the LMT Working Group as one of its communities of practice.

Since LMT's initial release, the Working Group, whose members include representation from HBO, Turner, Warner Bros, Disney, WWE, NBCUniversal, Paramount, Sony Pictures, Lionsgate, and Fox, among others, have reached agreement on:

- Language terms, definitions, and examples
- Template for adding new languages
- Over 200 language codes and display values -- with more on the way

---

[1] https://www.mesalliance.org/About-MESA

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

HBO encourages 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 LMT Working Committee.

Vendors and subject matter experts are also participating in the LMT development effort.

1.3 The LMT Mission Statement

The Language Metadata Table standard was created to provide a unified source of reference for language codes for 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 LMT

The LMT initiative began at HBO in 2017. Like many companies, HBO has multiple systems that are used across the enterprise. They include: production systems, marketing systems, scheduling systems and more. The metadata and taxonomy team were asked which code should be used to represent Latin American Spanish. Research showed that one system used *SPA-LA* while another one used *LAS,* and yet another one used *LATAM SPA*. And there were even more variations.

A project was created to normalize language codes across the enterprise and thus the Language Metadata Table (LMT) was developed. The initial table had 128 languages.

2.2 Approach: why IETF BCP-47?

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). IETF BCP 47 incorporates numerous ISO language and territory standards[3] in a precise combination, as well as allowable UN territory standards (UN M. 49) for those regions that have no ISO 3166 equivalent.

[3] http://www.iso.org/iso/home/standards/language_codes.htm

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[4], which resulted in the following findings:

- 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 value may be different from the Audio
    - Some languages are expressed differently, including spellings, ex: English, Chinese, Spanish
    - Audio languages may have multiple dialects dependent upon the geographic region

The outcome was to implement IETF BCP 47 language codes in order to achieve the required granularity of languages codes needed at HBO.

IETF BCP 47 is based upon existing metadata standards:

- ISO 639: Language codes
- ISO 3166: Country codes
- UN M. 49: UN Territory standards

IETF BCP 47 works because

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

IETF BCP 47 codes can be combined for greater descriptive granularity. So, for example, "English as spoken in the US" would be rendered "en-US", whereas English as spoken in Great Britain would be rendered "en-GB". This is helpful in accessibility applications, where it's important for the hearing-impaired user to understand the context of characters' speech, an aspect that can't be conveyed by mere closed-captioning - as well as in fully describing video content for the purposes of marketing and airing.

Unlike ISO 639, IETF BCP 47 was not developed as a bibliographic standard (though it has bibliographic applications). It was developed to describe languages within Internet applications and on the Web. Thus it is more suited to the purpose of describing the languages of digital assets than the ISO standard. It also provides for description of fictional languages (ex: Klingon), and so could handle HBO's Dothraki from Game of Thrones as well as other invented languages. IETF BCP 47 is an excellent solution to future-proofing the media and entertainment language metadata requirements.

[4] https://tools.ietf.org/html/bcp47#section-1

3. Common Language Metadata Needs Across Industries

Some standards have a single entity that needs to be coded for Language, but for media and entertainment, language codes are needed for:

- 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 values that can be applied for each language need.

4. LMT Language Grouping

Language groupings are an optional but 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 separate. A simple hierarchy allows for the maximum flexibility. Some language grouping examples are:

- Greek: to account for ancient vs modern
- English: British, Canadian, Australian, American, etc
- Spanish: Latin American vs European, Mexican vs Argentinian

Language Group Name	Language Group Tag	Language Group URN
Albanian	sq	urn:ietf:bcp:47:sq
Arabic	ar	urn:ietf:bcp:47:ar
Armenian Family	hyx	urn:ietf:bcp:47:hyx
Azerbaijani	az	urn:ietf:bcp:47:az
Basque Family	euq	urn:ietf:bcp:47:euq
Chinese	zh	urn:ietf:bcp:47:zh
Dutch	nl	urn:ietf:bcp:47:nl
Estonian	et	urn:ietf:bcp:47:et
French	fr	urn:ietf:bcp:47:fr
Greek	el	urn:ietf:bcp:47:el
Hindi	hi	urn:ietf:bcp:47:hi

Italian	it	urn:ietf:bcp:47:it
Japanese	ja	urn:ietf:bcp:47:ja
Latvian	lv	urn:ietf:bcp:47:lv
Malagasy	mg	urn:ietf:bcp:47:mg
Malay	ms	urn:ietf:bcp:47:ms
Mon-Khmer Languages	mkh	urn:ietf:bcp:47:mkh
Mongolian	mn	urn:ietf:bcp:47:mn
Pashto	ps	urn:ietf:bcp:47:ps
Persian	fa	urn:ietf:bcp:47:fa
Portuguese	pt	urn:ietf:bcp:47:pt
Serbo-Croatian	sh	urn:ietf:bcp:47:sh
Sign Languages	sgn	urn:ietf:bcp:47:sgn
Spanish	es	urn:ietf:bcp:47:es
Special*		
Swahili	sw	urn:ietf:bcp:47:sw
Uzbek	uz	urn:ietf:bcp:47:uz

Figure 1: LMT Language Grouping Table

*The Special grouping is for codes such as “und” (undetermined) and “zxx” (no linguistic content).

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 M.49, 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 - 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.

For LMT’s purposes, only the first three types of subtags are currently in use, though this may change with additional use cases.

Some examples of combinations:

Code	Language Description
zh-Hans	Chinese written in simplified script
nl-BE	Dutch as spoken in Belgium (Flemish)
ja-Jpan-JP	Japanese written with Han, Hiragana, and Katakana characters
sr-Latn	Serbian written in Latin script

Figure 2: LMT Language Combinations Example

IETF BCP 47 specifies that the shortest possible tag should be used. So if a language is being described without the context of the country in which it's spoken, only the primary (first) tag should be used. A use case for primary-only tags is in rights negotiations: a content company negotiating for the rights to distribute Spanish content isn't necessarily going to distinguish among the many types of Spanish that are spoken; in this instance, the simple code "es" would be used in systems that record this process. However, on the distribution side, it's very important to let trading partners and consumers know what sort of Spanish they'll be listening to or reading, so additional qualifiers would be necessary in those systems.

6. LMT Fields: Definitions and Examples

This section contains information about the LMT template, including definitions and best practices for population.

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 Armenian Family.
Language Group Tag	IETF BCP 47 tag.
Language Group Code	URN or URI for each language group 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 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 following IETF BCP 47 standard (which calls for the tags to be presented in Latin Script).
Visual Language Tag 2	Alternate script in which language is written following IETF BCP 47 standard (which calls for the 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 1 but not always.
URN	URN or URI for each language value in the LMT.

Figure 3: LMT Template Terms and Definitions

6.2 Populated Examples

This section contains examples of how the LMT is applied.

Column Header Name	**Example 1: English**	**Example 2: Spanish**	**Example 3: Serbian**	**Example 4: Mandarin**	**Example 5: Armenian (Eastern)**	**Example 6: Armenian (Western)**	**Example 7: American Sign Language**
Language Group Name	English	Spanish	Serbo-Croation	Chinese	Armenian Family	Armenian Family	
Language Group Tag	en	es	sh	zh	hyx	hyx	
Language Group Code	urn:ietf:bcp:47:en	urn:ietf:bcp:47:es	urn:ietf:bcp:47:sh	urn:ietf:bcp:47:zh	urn:ietf:bcp:47:hyx	urn:ietf:bcp:47:hyx	
Audio Language Tag	en	es-419	sr	cmn	hy	hyw	
Long Description 1	English	Spanish as Spoken in Latin	Serbian	Mandarin	Armenian	Armenian as spoken by the	American Sign

		America				Armenian Diaspora	Language
Long Description 2							
Audio Language Display Name 1	English	Español como se habla en América Latina	Srpski	普通话	արեւմտահայերէն	հայերեն	
Audio Language Display Name 2			cpncka				
Visual Language Tag 1	en	es-419	sr-Latn-RS	zh-Hans	hy	hyw	ase
Visual Language Tag 2			sr-Cyrl-RS				
Visual Language Display Name 1		Español como se habla en América Latina	Srpski	简体中文	արեւմտահայերէն	հայերեն	American Sign Language
Visual Language Display Name 2			cpncka				
URN	urn:ietf:bcp:47:en	urn:ietf:bcp:47:es-419	urn:ietf:bcp:47:sr urn:ietf:bcp:47:sr-Latn-RS urn:ietf:bcp:47:sr-Cyrl-RS urn:ietf:bcp:47:sr-RS	urn:ietf:bcp:47:cmn urn:ietf:bcp:47:zh-Hans	urn:ietf:bcp:47:hy urn:ietf:bcp:47:hy-AM	Urn:ietf:bcp:47:hy urn:ietf:bcp:47:hy-US	urn:ietf:bcp:47:ase

Figure 4: LMT Examples: Populated as Template Entries

7. Language Updates and/or Additions: Procedures

Any party can submit use cases to MESA's LMT working group, and these submissions will be considered periodically. The MESA working group will decide whether those use cases are incorporated into the LMT. The need for updates to existing languages or requests for new languages happens for a variety of reasons, as languages are in continual flux, including political.

When submitting an update or requesting a new language to be added, these are the steps to follow for when adding/updating languages in the LMT.

- Download the LMT template from the MESAlliance website
  - A new template needs to be populated for every language add/update
- Populate the template following IETF BCP 47 rules
- Submit to the LMT Working Group chairs for initial review: lmt@mesalliance.org
- The request(s) will be shared with the LMT working group for review and feedback.
- Once questions about the submission(s) are resolved, formal approval from the LMT working group will be requested.
- Upon approval, the changes will be added to the LMT and the updated table will be posted on the MESAlliance website.

- An email message describing the changes will be sent out. There will be two distribution lists:
    - Working Group: active working group members.
    - Implementers: contacts who need to work with/accept LMT values, but are unlikely to need to influence or change LMT.

8. Use Cases

This section contains definitions and examples of where and why languages, particularly for media and entertainment, need to be captured.

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), and there is a need to describe these with language metadata such as language type and script/writing system.

Use Case: to remain compliant with ADA and 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-cmn-Hans (Mandarin as rendered in Simplified characters)

8.4 Accessibility

There is an industry-wide need to be able to offer “visual description” - a narration of what occurs on the screen for the visually-impaired.

Use Case: to remain compliant with ADA and FCC requirements.

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

8.5 Acquisition/Rights

When acquiring rights to new content, organizations are not necessarily concerned with granularity of language data - rights are acquired to broadcast in Spanish, for example, rather than any particular flavor of Spanish.

Use Case: To describe an organization's right to distribute content in a specific language.

Ex: es (Spanish)

8.6 Trading Partners (Cable affiliates, electronic sell-through partners, etc.) and Consumer Viewing

Affiliates, trading partners, and consumers need to know the full description of the product they are receiving, including a granular description of language as it's spoken within the product. Therefore, the geo-specific tags would be used in those instances.

Use Case: To convey to consumers what specific language the product is in.

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

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		afi	British Sign Language
Norwegian	no		nb	Norwegian Bokmål

9. Examples

Example 1: Spanish

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

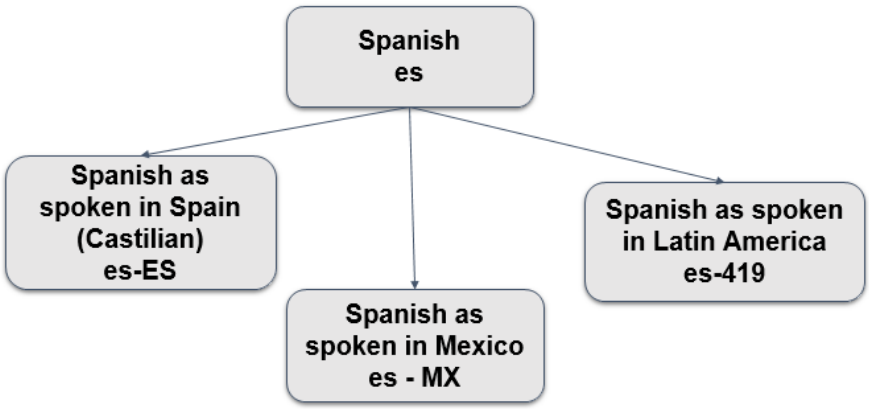

Figure 5: Spanish LMT Example: Generic, Latin America, Spain, and Mexico.

Example 2: Chinese

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

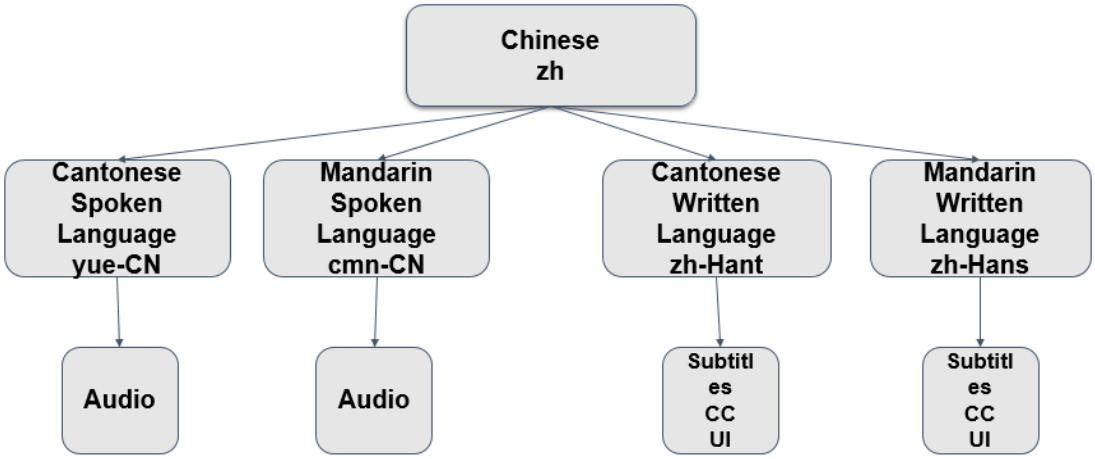

Figure 6: Chinese example for spoken and written languages.

Example 3: Italian/Neapolitan (*My Brilliant Friend*)

Below 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, so the LMT lists Italian and Neapolitan non-hierarchically.

Figure 7: Italian and Neapolitan Language Code Examples.

10. Compliance Best Practices

Compliance with LMT requires the use of LMT tags and/or URNs for languages covered by LMT use cases and that are included in the published LMT. That is, where LMT has defined one or more language tags and URNs for a use case, those codes and/or URNs 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 URNs that are not on the list.
- It is expected that those use cases will be submitted to LMT as soon as practical.

- However, if LMT has rejected a use case, those tags and URNs should not be used.

It is also possible that existing and/or legacy systems/applications may not be able to comply with LMT codes due to a variety of reasons. If those systems can't 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.

In addition, 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 HBO experiences.

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

Additionally, IETF BCP 47 provides for distinctions between spoken and written language, with the "Script" subtag. This allows us 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 committee will continue to meet on a monthly basis to discuss policy and procedures, and to make necessary changes as needed.

Appendices

Appendix 1 - Versions of the LMT

Here are the list of links to the current LMT documentation.

Version	**Link**	**Format**
LMT v2.0	(TO COME) Will be posted on the MESA website.	Excel and PDF
LMT Template	(TO COME) Will be posted on the MESA website.	Excel
Policies and Best Practices	(TO COME) Will be posted on the MESA website.	PDF
LMT Overview Presentation	(TO COME) Will be posted on the MESA website.	PDF
LMT v1.0 (DEPRECATED)	https://www.mesalliance.org/wp-content/uploads/2018/07/LMT-v1.0-07-31-18.pdf	PDF Download
LMT v1.0 (DEPRECATED)	https://www.mesalliance.org/wp-content/uploads/2018/07/LMT-v1.0-07-31-18.xlsx	Excel Download

Appendix 2 - Resources

Here is the list of resources and references for IETF BCP 47 and LMT.

IETF BCP 47 available at: https://tools.ietf.org/html/bcp47#section-1
IETF BCP 47 Complete Language Registry. Available at: http://www.iana.org/assignments/language-subtag-registry/language-subtag-registry
ISO 639 available at: http://www.iso.org/iso/home/standards/language_codes.htm

ISO 639-2: ISO 639-1 Mapping available at: https://www.loc.gov/standards/iso639-2/php/code_list.php
UN M.49 Codes available at: https://unstats.un.org/unsd/methodology/m49/
MESA Alliance Articles: https://www.mesalliance.org/language-metadata-table https://www.mesalliance.org/2018/08/08/hbo-looks-to-demystify-language-metadata/ https://www.mesalliance.org/2018/08/07/mesa-publishes-hbo-developed-me-industry-language-metadata-table/
About MESA page: https://www.mesalliance.org/About-MESA

Appendix 3 - Contact List

This table contains the contact information for key LMT and MESAlliance resources.

Contact Name(s)	Email Address	Description
Yonah Levenson, Laura Dawson	LMTChairs@mesalliance.org	LMT Co-chairs
LMT Working Group	LMTWG@mesalliance.org	Includes those who are active members of the LMT Working Group
General LMT Info	LMT@mesalliance.org	Use this email address when requesting general information and/or administrative requests about LMT.
MESAlliance Info	https://www.mesalliance.org/contact-us/	Use this email address when making general queries about MESAlliance.

LMT working group contributing companies include, but are not restricted to:

- Digital Bedrock
- Discovery
- Disney
- EIDR
- Fox
- Gracenote
- HBO
- Hasbro
- Lionsgate
- Movielabs
- NBCUniversal
- Paramount
- Showtime
- Sony Pictures
- Turner Sports Library
- Warner Bros
- WWE
- And more.