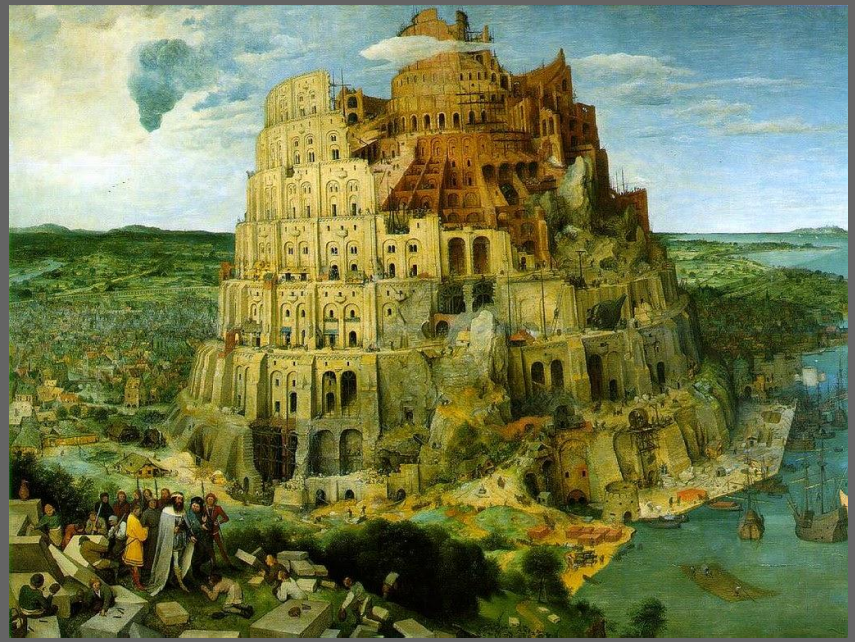
Yonah Levenson, WarnerMedia

Language Metadata Table Co-Chair
March 16, 2021

LANGUAGE METADATA TABLE

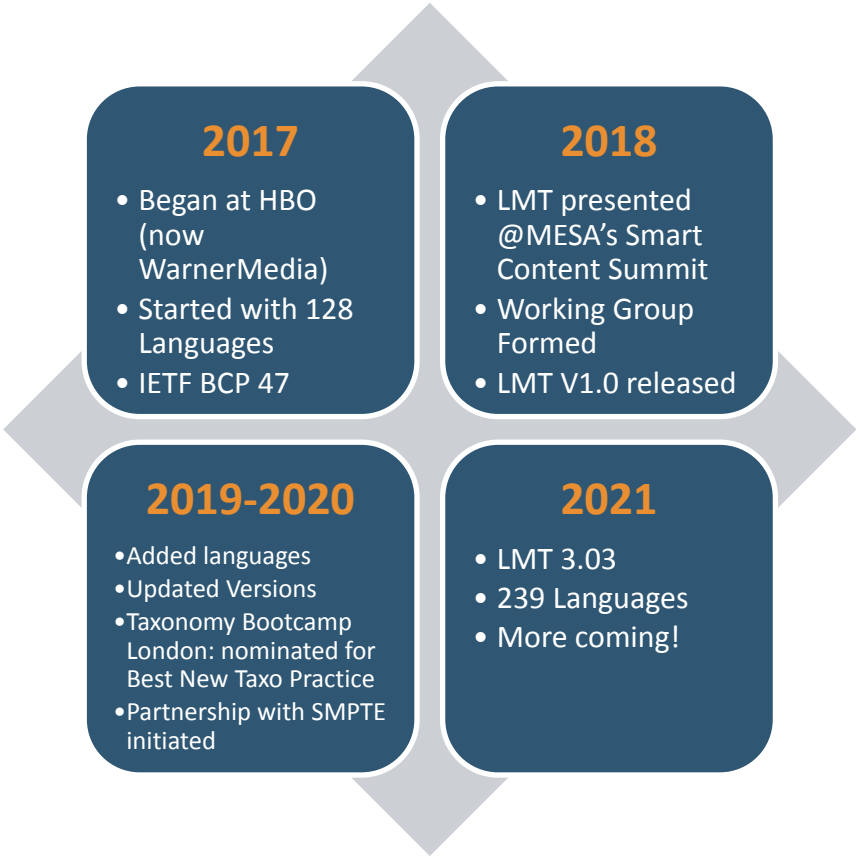
No Longer the Tower of Babel



ENDORSE.

THE EUROPEAN DATA CONFERENCE ON REFERENCE DATA AND SEMANTICS

LMT:
Then and Now



The LMT Mission Statement

*The Language Metadata Table (LMT) was created to provide a unified source of reference for language codes for use throughout the media and entertainment industries**

**(it works for others too!)*

MESA SMPTE LMT Working Group Members & Contributors

Advantages of Adopting LMT

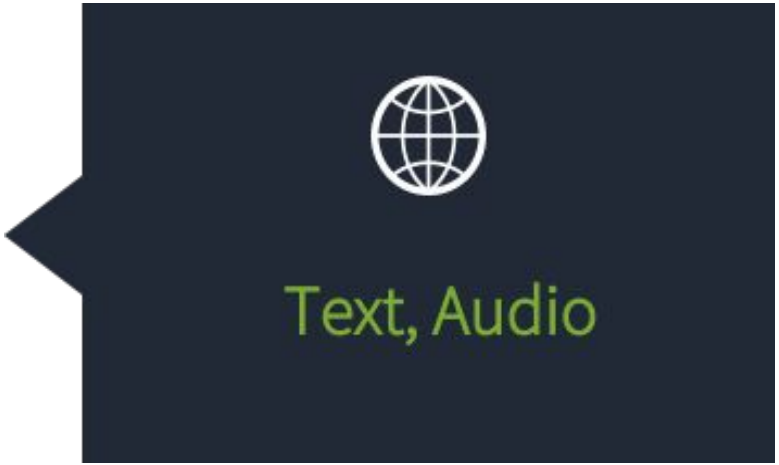
The Working group: Checks and balances across the industry

Standardized distinctions between spoken and written languages

Consistent codes between service providers, clients, and content owners

LMT Scope: Populate asset language elements (text, audio)

- Languages only
- Audio
- Notation of script/writing system included
- *Endonyms:* Language name in the country's language (Français) as well as English (French)



LMT Use Cases Include:

1. Audio
2. Visual (Written/Text)
   - Subtitles
   - Closed Captions
   - Burned in/Forced Narratives
   - Accessibility (SDHH, Sign Language)
3. Content licensing
4. Rights
5. Acquisition
6. Distributing content
   - One or more languages
   - Electronic Sell Through (EST)
7. Localization preferences

IETF BCP 47: Use the Shortest Code

- IETF: Internet Engineering Task Force (a.k.a, the Internet people)
- BCP: Best Current Practice
- BCP 47: Tags for Identifying Language

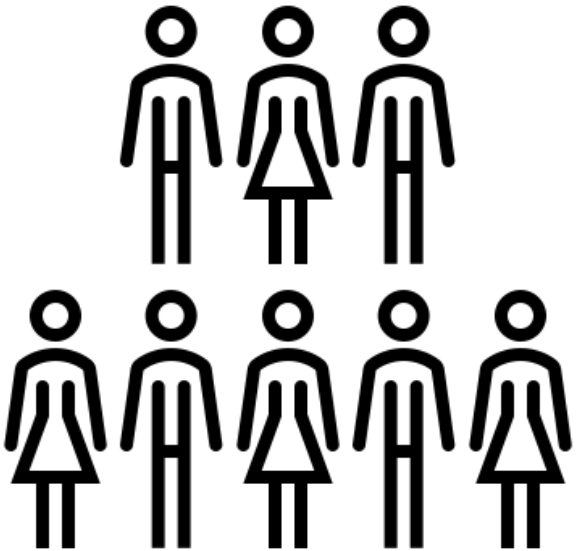
Standard	Pattern
ISO 639	2 and 3 character Language Codes
ISO 3166	2 character Country Codes
UN M. 49	3 digit numeric Territory Codes
ISO 15924	4 character Script Codes

- Over 40,000 possible combinations
  - **General:** fr for undifferentiated French
  - **Specific:** fr-FR for French as spoken in France vs fr-CA for French as spoken in Canada (Quebecoise)
- A W3C Standard

LMT Examples

Column Header Name	English	Spanish	Serbian	Mandarin	Armenian (Eastern)	Armenian (Western)	American Sign Language
Language Group Name	English	Spanish	Serbo-Croatian	Chinese	Armenian Family	Armenian Family	
Language Group Tag	en	es	sh	zh	hyx	hyx	
Audio Language Tag	en	es-419	sr	cmn	hy	hyw	
Long Description 1	English	Spanish as Spoken in Latin America	Serbian	Mandarin	Armenian	Armenian as spoken by the Armenian Diaspora	American Sign 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			српски				
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			српски				

LMT Working Group: Current State



- Requests for 150+ additional languages!
  - Sources include: WarnerMedia, ViacomCBS, Disney
  - MESA resource is working through the requests and suggesting codes
    - Language Groups
    - Individual Languages
- Chinese
  - Additional Chinese dialects (Audio) have been submitted for review and approval
  - Request to separate Visual and Audio languages/dialects is under consideration
- Languages of the Indian Sub-continent
  - Qube Cinema is working on the Language Group Codes and Language Codes
  - Rolling submissions for approval

LMT Exists Because …

MESA

Media and Entertainment Support Association

- **Supports** service providers in advancing efficiencies in the creation, production, and distribution of media and entertainment
- **Promotes** thought leadership, industry initiatives and accomplishments of our members and their customers
- **Connects** all with a culture of community

- **S**ociety of **M**otion **P**ictures & **T**elevision **E**ngineers
- Internationally recognized Standards organization
- More than 800 engineering standards
- Developed in a collaborative process with individuals and corporations to advance global interoperability of hardware and software
- Results: improved workflow and uncompromising quality for seamless creation, management and delivery of media

LMT Sponsorship/Support and Technical Home

MESA

- Adoption Guidelines
- Templates
- Resources
- Source for published materials

- Tools
- Infrastructure
- Formalized Standards Process
- Validator

Additional Standards Partners include:

EIDR: Entertainment ID Registry

ISDCF: InterSociety Digital Cinema Forum

MovieLabs

10E	20F	21DC	24TB	25CSS	30MR	31FS	32NF	34CS	35PM	Misc
Metadata & Registers Technology Committee										

LMT is in SMPTE's Technical Committee 30 [of the] Metadata Registry

Bruce Devlin,
SMPTE Standards VP
svp@smpte.org

- **Tech Committee Scope**
  - Definition and implementation of the SMPTE Registration Authority, used to identify digital assets and associated metadata.
  - Common definition of metadata semantic meaning across multiple committees.
- **Status**
  - Currently in the project approval stage (10% done)
  - Validator available for preview

- **Topics and Standards**
  - Metadata Registers
  - Vocabularies for other TCs
  - https://smpte-ra.org
  - UMIDs

Resources and Links

- General Inquiries: LMT@mesaonline.org
  - Yonah Levenson (WarnerMedia) and Meg Morrissey (Netflix): co-chairs
- LMT Codes in Workbook and XML formats, plus documentation and templates: https://www.mesaonline.org/language-metadata-table
- SMPTE Validation Tool Demo https://drive.google.com/file/d/164EiH9DwvR7pDusKp7hVv5PLW4d7aYk7/view?usp=drivesdk

Primary LMT Code Validation Sources

- Validator for checking codes: https://r12a.github.io/app-subtags/
- ISO 639-3 Registration Authority: https://iso639-3.sil.org/
- IANA Language Subtag Registry https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry
- And others

Thank You!

Yonah Levenson: LMT@mesaonline.org
@yonahleve yonah.Levenson@warnermedia.com

ENDORSE.
THE EUROPEAN DATA CONFERENCE ON REFERENCE DATA AND SEMANTICS

Publications Office of the European Union

ISA²

Language Metadata Table

https://www.mesaonline.org/language-metadata-table

APPENDIX

The LMT Mission Statement

The Language Metadata Table (LMT) was created to provide a unified source of reference for language codes for use throughout the media and entertainment industries.

- To create a standardized table of language codes for implementation by entertainment and other industries using IETF BCP 47 (a.k.a., RFC 5646).
- 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.

Anatomy of a Language Code

- Full code syntax: **language-script-region-variant-extension-privateuse**
  - e.g., **mn-Cyrl-MN** for Mongolian written in Cyrillic as used in Mongolia
- Selecting from 9,000 subtags to create 40,000 combinations is overwhelming
- LMT: pre-constructed codes supported by use cases
- Language groupings are explicitly defined – easy enough for Spanish, but hard for Chinese
- For each language, several fields are used to identify the standard:
  - Language Group Name, Tag, Code
  - Audio language tags and displays
  - Visual language tags and displays
  - Descriptions

LMT Metadata Field Names & 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 or URL for each language value in the LMT.