The language metadata table: Providing a single, unified language reference source for media and entertainment

Received (in revised form): 25th February, 2022

Yonah Levenson

Co-Director, Digital Asset Management Certificate Program, Rutgers University, USA

Yonah Levenson is the Co-Founder and Co-Director of the Digital Asset Management Certificate Program at Rutgers University, and Rutgers Professional Development Instructor of the Year for 2020–2021. She is also the founder and Co-Chair of the Language Metadata Table, an industry standard for language codes. Her previous positions include Vice President/Group Director of Information Services at Ad-ID, Manager of Metadata Strategy and Taxonomy Governance at WarnerMedia and HBO, and Senior Metadata Analyst at Pearson Education.

Rutgers, The State University of NJ, School of Communication & Information,
4 Huntington Street, New Brunswick, NJ 08901, USA
E-mail: yonah.levenson@gmail.com

Bruce Devlin

Founder, Mr MXF, UK

Bruce Devlin is Founder of Mr MXF and past Standards Vice President for the Society for Motion Pictures and Television Engineers (SMPTE). Over the last 30 years, Bruce has designed everything from application-specific integrated circuits to algorithms, and has become well known in the media industry for his work on files and systems, particularly MXF and IMF. Bruce is an alumnus of Queens' College Cambridge, fellow of the SMPTE as well as the recipient of SMPTE's David Sarnoff Medal and the International Moving Image Society Achievement Award.

Mr MXF, First Floor Ridgeland House, 15 Carfax, Horsham, West Sussex, United Kingdom, RH12 1DY
E-mail: bruce@mrmxf.com

Eric Emeric

Senior Director of Metadata Management & Data Governance, Showtime Networks, USA

Eric Emeric is the Director of Metadata Management & Data Governance at Showtime Networks, where he leads all global metadata and data governance endeavours related to customer-facing data flows within enterprise systems, from inception to distribution. He has been a data professional for the last 23 years, establishing metadata best practices in the medical, legal and entertainment fields. Eric has a master of library science from Queens College in New York, and currently is a virtual volunteer for the Library of Congress.

Showtime Networks Inc, 1633 Broadway, 15th Floor, New York, NY 10019, USA
Tel: +1 212 708 1206; E-mail: eric.emeric@showtime.com

Sarah Nix
Senior Director of Archives and Data Governance, Paramount Global, USA

Sarah Nix is Senior Director of Archives and Data Governance at Paramount Global, where she is responsible for a collection of over 2 million physical and digital assets as well as the metadata used to uniquely identify them. As part of her role, she sits on the board for EIDR, the universal identifier registry for movie and television assets. She is also the Co-Chair of Women+, which supports the strong community of women within the company. Sarah holds a BS in communications from Ithaca College's Roy H. Park School of Communications, where she serves as on the Alumni Association Board of Directors.

Paramount Global, 1515 Broadway, 17th floor, New York, NY 10036, USA
Tel: +1 212 846 8270; E-mail: sarah.nix@viacomcbs.com

Abstract The Language Metadata Table (LMT) offers a single source of reference for language codes for use throughout the media and entertainment ecosystem. This paper explains how it came to be, and why it has been embraced so enthusiastically. The paper also discusses the role of the Society of Motion Picture and Television Professionals (SMPTE) as the standards body responsible for the provision of language codes adopted by the LMT. This is followed by a case study of how Paramount Global has leveraged the LMT to resolve its language needs following a series of mergers and acquisitions.

KEYWORDS: language, metadata, media and entertainment, content distribution, terminology, localisation

INTRODUCTION

Initially developed by WarnerMedia's HBO and launched in mid-2018, the Language Metadata Table (LMT) was created to provide the media and entertainment industry with a single, unified standard of language terminology. An expandable mapping resource — essentially, a human-readable reference table — the LMT organises language metadata for more than 271 language codes and display values (with roughly 30 more currently being researched) to provide data specialists with a single, open source table of language metadata values. The LMT includes codes for audio and timed text for content; rights and licensing localisation; distribution territories; and accessibility for the visually and hearing impaired. Its uses are vast, covering standardised distinctions between spoken and written languages, the licensing of international content, distribution of non-English content and end-user localisation preferences.

In 2016 Yonah Levenson, then HBO's newly hired Manager of Metadata Management and Taxonomy, attended a company product development kickoff meeting. Little did she know that meeting would end up influencing how the entire media and entertainment industry thinks about language codes.

One developer asked during the meeting what the language code for Latin American Spanish should be. He had looked through various internal systems and not one system had the same code — 'LAS', 'LATAMSPAN', SPA-LA', etc. It struck Levenson and colleague Laura Dawson, then HBO's metadata analyst, that something needed to be done to rectify this. They went to work researching language codes across HBO's systems and discovered the many languages had multiple language codes.

The Latin American Spanish example is indicative of a common problem when coding languages: there has been no preferred way of handling geographic regions

instead of countries and/or states. When it comes to language codes, organisations tend to look to ISO 639 — the set of standards covering the representation of names for languages and language groups. ISO 639 is often combined with ISO 3166, which contains country codes. However, ISO 3166 does not cover geographic regions.

Further research led to the discovery of IETF BCP 47[1] — a language standard published by the World Wide Web Consortium. This in turn led to the creation of the Language Metadata Table (LMT).[2]

IETF BCP 47 (aka RFC 5646) takes advantage of both ISO 639 and ISO 3166, as well as a third standard: UN M.49, which specifies unique three-digit numeric codes for different territories. This makes it possible for IETF BCP 47 to include a geographic territory in the language codes it prescribes. Thus, instead of having multiple makeshift codes for Spanish as spoken in Latin America, the valid code becomes 'es-419' (the code for Spanish is 'es' with the code for Latin America in UN M.49 being '419').

Why is the LMT needed when IETF BCP 47 accommodates languages and territories, in addition to country codes? The issue is the possible number of code combinations — more than 40,000 — that can be created with IETF BCP 47. This means that while an organisation can justifiably claim to be BCP 47 compliant, the source language code and the target code can be vastly different.

SOLVING AN UNACKNOWLEDGED PROBLEM

After that initial 2016 meeting, Levenson decided that the language codes used across HBO needed to be implemented and applied consistently. The benefits were quickly made apparent when the metadata fields and codes were standardised across systems. Accurately populated metadata fields led to less confusion across departments and reduced costs.

Research identified various areas where language codes would be beneficial, including:

- *Audio:* to provide anyone working with the assets, including vendors, clients, translators, marketing, etc with a standardised description of each of the audio languages associated with the content;
- *Closed captions and subtitles:* to distinguish between (a) the audio language associated with a piece of content and (b) the caption or written language when sending materials to vendors;
- *Burned in or forced narratives:* physical signs in content often appear in a written language that is different from the subtitles associated with the audio language;
- *Accessibility:* to distinguish visual descriptions and/or American Sign Language from other languages used in the content;
- *Acquisition/rights:* the language(s) included in agreements regarding territorial or distribution rights for the content;
- *Electronic sell-through (EST) partners:* to display languages onscreen in the correct dialect with respect to the subtitle or audio languages.

As the LMT had to be sufficiently detailed to satisfy all of these use cases, names were developed for the specific fields of data required for each language. Studios release content internationally, hence it was necessary for the LMT to include the endonym, or how the language is referred to natively.

Sixteen months later, 127 languages were accounted for in the initial internal release of the LMT. Today there are more than 271.

System development typically involves representation from multiple departments, including system sponsors, data architects, developers, subject matter experts and end users. The metadata field names to include in the system are agreed between all involved resources. Spreadsheets are a good tool for capturing existing and proposed metadata

field names because, once populated with the field names from each system, it is fairly straightforward to identify commonalities and differences. The analysis should also include standardised metadata field names and lists of values. When the field names and controlled vocabulary values differ between systems, mapping tables are necessary to be sure the values in the fields transfer and convert properly between systems.

It is best practice to apply standardised metadata field names and controlled vocabulary values whenever feasible. When standards exist and are implemented, the need for mapping tables is reduced or eliminated, which makes quality assurance checks less time-consuming. Consistent field names and values also provide for an overall better client experience.

IETF BCP 47 is an extremely flexible metadata standard for language codes. There are many types of languages, and IETF BCP 47 accommodates them all. Examples of different types of language include:

- *audio:* the language that one hears — examples include Chinese dialects such as Hakka and Hunanese;
- *script:* the writing system that one reads — examples include traditional Chinese and simplified Chinese, where the script or written language differs from the spoken dialect;
- *visual:* the communication that one sees — for example, sign language.

IETF BCP 47 works because:

- language, dialect, script and geographic codes can be combined in more than 40,000 ways;
- it covers everything from the general (en for English) to the specific (fr-FR versus fr-CA, to distinguish Parisian French from Quebecois); and
- codes are under regular review to ensure they remain current (eg 'Greenlandic' updated to 'Kalaallisut') to reflect contemporary cultural norms.

The following fields and standards can be applied to an IETF BCP 47 language code (note that the LMT uses values from the first three bulleted standards only):

- *ISO 639:* two- and three-character language codes;
- *ISO 3166:* two-character country codes;
- *UN M.49:* three-digit numeric territory codes; and
- *ISO 15924:* four-character script codes.

The full code syntax looks like this:

language-script-region-variant-extension-privateuse

For example:

mn-Cyrl-MN

represents Mongolian written in Cyrillic as used in Mongolia.

THE MAKING OF LMT

It was initially expected that it would take about three months to make the language codes consistent across the enterprise. In reality, however, language codes were found to reside in far more systems than anticipated, and the usages were more complex than expected. Resolving this took more than a year.

The team had to work through complex use cases, including when and where to apply audio language codes and written language codes, depending on the language. Arabic, for example, proved just as challenging as Chinese when it came to identifying the differences between audio/dialects and script codes.

In addition, languages spoken in multiple countries, such as English, Spanish and French, need to be organised into groups. As seen in Figure 1, the team at HBO analysed how and where language codes were used and developed a template containing the types of information that must be described for each and every language in the LMT. Figure 2 shows examples of completed LMT records for various languages.

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.

Figure 1: Definitions for each of the field columns included in the LMT

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	Western Armenian	American Sign Language
Long Description 2							
Audio Language Display Name 1	English	Español	srpski	普通话	հայերեն	արեւմտահայերէն	
Audio Language Display Name 2			српски	国语			
Visual Language Tag 1	en	es-419	sr-Latn	zh-Hans	hy	hyw	ase
Visual Language Tag 2			sr-Cyrl	zh-Hant			
Visual Language Display Name 1	English	Español	srpskohrvatski	简体中文	հայերեն	արեւմտահայերէն	American Sign Language
Visual Language Display Name 2			српскохрватски	繁体中文			

Figure 2: Examples of multiple languages and the codes that have been applied to each

THE PATH TO STANDARDISATION

In late 2017, Levenson and Dawson presented a case study of the LMT at a Media and Entertainment Services Alliance (MESA) meeting. At that time, studios and content distributors were struggling with the application of inconsistent language codes. When they saw the results that HBO had realised, however, they recognised that they were looking at a solution to media and entertainment's language coding problems, and adopted LMT as an industry standard.

The Academy Award winning movie, 'Roma', provides a great use case of what happens when standardised language codes are not provided when the creation of subtitles is requested. The director, Alfonso Cuarón, was angered that the subtitles created for the Spanish release of the movie were in Castilian rather than the native Mexican-Spanish dialect that is spoken in the movie.[3] Netflix had to remove the movie from Spain until the subtitling issue was corrected. Ideally, the work order for the subtitles should have included the correct LMT code, and this thorny problem (which probably cost Netflix considerable money) would not have happened.

The need for standardised language codes led to the formation of the LMT Working Group in 2018, sponsored by MESA. Since then, the working group has researched, documented and approved additional languages beyond the original 127. As of early 2022, the current count of approved languages is up to 271. Another 30 languages are still being researched and will be presented to the group for review, followed by a vote for approval.

Research is typically conducted by subject matter experts who are familiar with the languages that have been submitted for consideration. Recent subcommittees include: Arabic (with subject matter experts from the Library of Congress, Ryerson University, Al Jazeera and Gracenote); languages of the Indian Subcontinent (led by Qube Cinema); and Chinese and other Asian languages. Research continues for additional languages including Quechua (Peru), Native American languages, and more.

SUPPORT FROM THE MEDIA AND ENTERTAINMENT COMMUNITY

Participants in the LMT working group typically come from other standards bodies, studios and other organisations in the media and entertainment space, and include WarnerMedia, Disney, Sony, APEX (Airline Passenger Experience), Qube Cinema, MovieLabs, Gracenote, EIDR (Entertainment ID Registry), Lionsgate and Paramount (formerly ViacomCBS).

In correspondence with the authors, Nona Jansen Walls, of industry leader Slalom Consulting, commented:

> 'We all know by now that clean and consistent data are a key factor to reduce friction in the media supply chain and enables production and distribution at scale. Adopting MESA's LMT standard for spoken and written word provides that consistency, from pitch to platform, and eliminates guesswork when reaching out to localised and global audiences alike.
>
> LMT is the only language standard developed for our industry's unique needs. It's been developed and tested with industry experts across production companies, studios, and distributors, and can organise and make sense of "language" among scripts, subtitles, dubs, metadata localisation, credits, rights licensing, analytics, and more. Using any other "standard" (or no standard at all) is a ticket to endless data mapping and reconciliation, and who wants to do that?'

Likewise, in similarly private correspondence, Andy Beer, VP of Technical Engineering at Spafax Hub, added:

> 'The in-flight entertainment ecosystem operates with abundant provision for regional language requirements. Setting consistent code points for language assignment (content metadata) has been largely impossible so far. The LMT project

> is the best hope for standardisation. APEX.aero (the Airline Passenger Experience Association) will be recommending the universal adoption of LMT to simplify the in-flight entertainment supply chain'.

As interest in LMT has continued to increase, several standards bodies have expressed their interest in including and incorporating LMT in their standard or suite of standards. The LMT chairs and MESA recognise that adoption would increase if there was a way to programmatically validate the LMT codes.

A DEEPER, TECHNICAL DIVE

As every language tag within the LMT is based upon IETF BCP-47, it is important to understand what this means.

IETF BCP-47[4] provides the rules for the formatting of tags and sub-tags. It also appoints the Internet Assigned Numbers Authority (IANA) as the registrar responsible for listing all the valid tags. It is important to remember that these tags are dynamic. Countries and languages change over time, and as codes are assigned or withdrawn by ISO 639, ISO 15924, ISO 3166 and UN M.49, the person responsible for maintaining the language register must evaluate each change and determine the appropriate course of action according to the rules in IETF BCP-47.

To this end, IANA[5] maintains a list of the various language codes that might exist in the world at any given time. The role of the LMT is to enumerate the subset of those codes that is in active use within the professional media and entertainment space. The LMT committee could simply put a list online and reference that, but if the committee goes through a standardisation process, this enables other entities (large companies, studios and broadcasters, government entities) to reference the LMT, knowing that the list is being maintained and that the values it contains can be trusted for interchanging content within the professional media space.

To standardise the LMT, the committee must therefore be able to reference an existing accredited standard for the tags (ie IETF BCP 47) and then to follow a due process to publish the list of tags in use. The Society for Motion Picture Television Engineers (SMPTE), (2022), available at https://www.smpte.org/ was chosen as the standards body to conduct this work and already registers a certain number of controlled lists.

SMPTE DOMAIN USAGE

The role of the standard is to enumerate a subset of the IANA tags. As the tags will change over time, it therefore must be recognised that the values in the LMT will also change over time and may be used in environments where a single piece of content may need multiple tags. There may also be interesting business logic — outside the scope of the standard — that encourages process automation using the tags as controlling metadata. The goal is to keep the standard focused on having a controlled vocabulary of labels with clear definitions and a limited amount of extra metadata associated with each entry. It is useful to standardise some of these additional metadata and to leave everything else up to the application or device. In determining what gets standardised in the register, it helps to pose some questions that the registered values can be used to answer. For example:

- What languages are spoken in this clip? For example, Castilian
- What languages are spoken by the target audience of this clip? For example, Castilian — es-ES
- What languages are written in this clip? For example, French
- Are there alternative names for this language? For example, Español, Castellano, or for French — Français
- What languages does my facility require for broadcasting *and* distributing a clip?

Ideally, a standard should contain just enough data to allow a tag to be used interoperably in any of these situations without requiring duplication. The standard should also not be too burdensome on the end user. This means that someone inserting the tags at source may know that the content is in Spanish, but they may not be expert enough to know the difference between the Spanish spoken in Spain or Mexico or Chile. The compromise within the LMT is that there are standardised tags for spoken language, visual language and language groups.

A user who does not know which specific version of Spanish is being used might use a group code to indicate that it is from the general Spanish set of languages.

Professional media companies are likely to label their content depending on how that label is being used. A production company might label a title with the tag to indicate that they intend this version to be shown to a Spanish-speaking audience. A distribution company, however, might re-label the same file with the codes es-ES, es-AR and es-BO to indicate that the Spanish content has been quality-controlled for television for Spanish speakers in Spain, Argentina and Bolivia. A different distribution company handling cinema releases might have a different labelling system for their specific business case.

A due process standards body, like SMPTE, may create additional metadata to help differentiate these cases if the volunteer members of the controlling committees deem this to be necessary.

THE (OPEN) SMPTE PROCESS

The SMPTE process[6] is open to anyone. This means that anyone can become a member of the committees and contribute to, or vote on, any document being standardised. Every document starts life as a project with a defined scope of activities that is approved by the members of the SMPTE committees. A drafting group is formed to create a control document that defines the semantics and syntax of the register. It also defines the process for updating and maintaining the register, building upon the precedent for other registers maintained by SMPTE. This document (a 'committee draft') is then made available on the SMPTE website and GitHub for public comment and scrutiny. When the committee feels that sufficient time has elapsed, then the document is voted upon, and the structure of the register becomes a full SMPTE standard.

WHAT IS BEING STANDARDISED?

Elements of the register are proposed to the committee and added according to the procedures in the control document. Simply put, every new addition must be voted upon to show that it has been scrutinised by the committee. SMPTE will then publish the tags in an appropriate fashion at https://smpte-ra.org. The current plan is to use an open source tool contributed by a SMPTE member to manage the register. This provides human and machine viewing tools, as well as conversion functions, to import and export with ontology management tools and differencing functions to see what has changed between versions of the registers.

The proponents of the LMT decided that a JSON[7] (JavaScript Object Notation) format was the preferred format for interchanging the LMT between systems rather than managing the ontology. The SMPTE process is therefore oriented towards maintaining the interchange representation of the register, a proposed snapshot of which is shown in Figure 3. Users of the LMT may store the register in a different format that is optimised for their management, business or technical needs. This is outside the scope of SMPTE's work.

JSON is a very simple representation of Objects that have properties. Readers familiar with XML should be aware that the goal of XML is to mark-up documents to allow the exchange of data with associated rich metadata in a consistent format. The

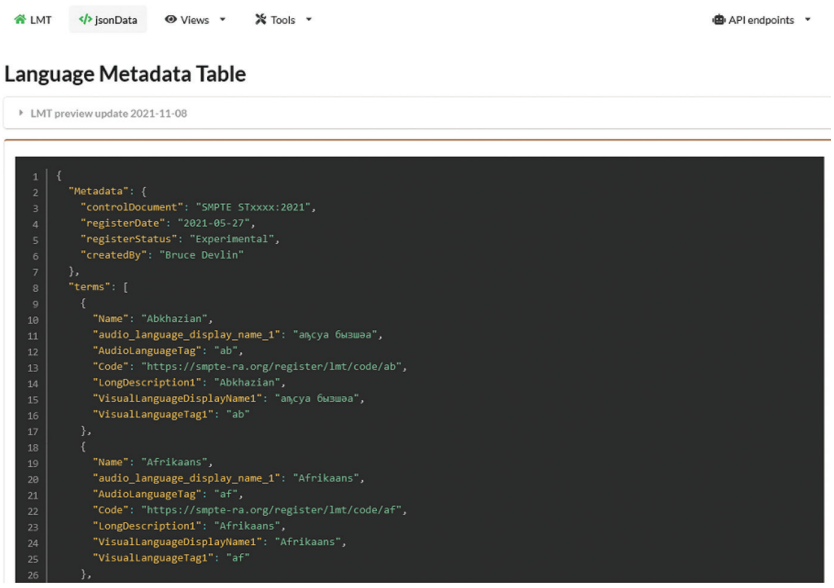

Figure 3: JSON Registry: A snapshot from the prototype tool showing a portion of the JSON register submitted to SMPTE

goal of JSON is much simpler — to allow the exchange of Javascript Objects via text documents. The JSON syntax is very simple and visually easy to read. It does not allow comments because these have traditionally been misused in HTML and XML to provide programmatic workarounds such as providing secret instructions to XML or HTML parsers (pragma statements). JSON's goal is very simple — to exchange objects. Just like XML, the semantics or permitted syntax of any specific JSON document must be provided via external means such as a schema, PDF document or some other instruction for a human or a machine to interpret.

Objects and properties have names (keys) and may be grouped into arrays. The SMPTE LMT JSON is structured to make it relevant to users of the data to present it, to interchange it or to manipulate it within a product rather than for ease of management. A schema is provided to check that the core syntactic rules are followed. An example of the schema is shown in Figure 4. It is expected that a common usage of the LMT will simply be to display the JSON's Audio Language Tag field. To disambiguate this field, the JSON has a code field of the form 'https://smpte-ra.org/register/lmt/code/af'. It is SMPTE's goal that this should resolve in a browser to a human readable view of the register entry.

PROOF OF LMT'S IMPORTANCE

Every organisation that creates, distributes and/or sells content to third-party entities benefits from having an established data governance programme in place. Effective data governance — the process of managing the availability, usability, integrity and security of data in enterprise systems, based on data standards and polices that also control data usage — ensures that data are consistent and trustworthy and do not get misused.

The advantages of effective data governance cannot be understated: it helps avoid inconsistent data lanes in different departments and business units; agreements emerge around data definitions, allowing for widespread and shared understanding of the data; and data

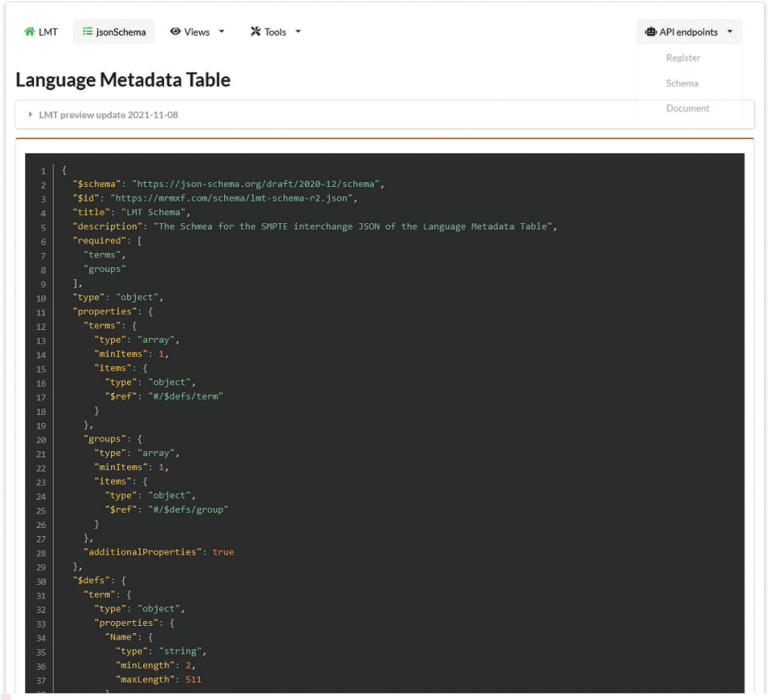

Figure 4: The JSON schema is created to ensure the syntactical accuracy of the JSON that is generated; it does not check for duplicate entries or other second-level errors

quality improves via efforts to identify and fix errors in data sets high in the supply chain.

The increase of analytic accuracy, providing decision-makers with reliable information, the implementation and enforcement of policies by committee to help prevent data errors and misuse, and the removal of duplicative data workflow between business units, all emerge with effective data governance.

Further, the effective use of what the LMT provides fits perfectly in any media and entertainment firm's data governance plan.

A MAJOR USE CASE

Viacom began its data governance strategy in 2015. A solution was needed to mitigate inconsistent data, particularly in reporting to the company's research and advertising sales teams.

The biggest hurdle was not technology but change management within the organisation, resulting in the creation of a centralised team to perform the governance task companywide. There were too many different teams managing metadata in siloed workflows and systems. To be successful with the governance implementation, Viacom needed to work with these individual teams and demonstrate the benefit of having a central location for the metadata, with a tighter focus on valid data entry. Reducing duplicative data entry by having this metadata flow into the downstream systems was another benefit to the approach. Within the governance workflows, the company implemented a two-person validation, which has one team member performing the first review and entry and a second team member checking that work for accuracy.

As Viacom determined which fields would be most beneficial, and which standards to embrace, a decision was made to adopt what the LMT uses as the language standard, namely IETF BCP-47. Relying on industry standards reduced any ambiguity in the implementation of governance strategy. If there is a standard that has been widely adopted and used among

peers, it holds more weight with internal groups who might be hesitant.

In 2018, then owned by CBS, Showtime Networks began its data governance strategy. Fortunately, all major stakeholders were eager for data structure and were able to lean into the data governance process seamlessly. Shortly thereafter, Showtime joined the LMT working group, keen on the group's goal to create a language standard that the entire content producing industry could follow. Here, having a data governance programme already in place was a major advantage, as all system stakeholders were readily available.

Once approval was obtained from Showtime's data governance board to add LMT to the project list, the next task was to collect all language tables used by internal enterprise systems. Once all tables were collected, a matrix cross-referencing the various language tables in variant standards across all internal systems became the genesis of what would later become Showtime's standardised language metadata table.

Once the companies merged to create Paramount, Sarah Nix, Senior Director, Archives and Data Governance, and Eric Emeric, Senior Director, Metadata Management and Data Governance, joined forces to investigate the best path to implement the LMT across the entire company. Just as was done at Showtime, all language tables across Paramount's systems were collected and checked for omissions, with collaboration across networks making it possible to fill any gaps in the enterprise systems with either legacy Viacom or CBS codes.

After identifying and gathering the various language tables from across Paramount, many Zoom sessions were spent aligning the LMT against both Showtime and legacy Viacom language codes in endless rows in Excel.

Once all data had been captured in an ever-growing Excel spreadsheet, a detailed comprehensive audit was necessary. A line-by-line audit comparing LMT codes against all Showtime and Viacom language codes was conducted. Although it took a few weeks, this auditing step was necessary to gain an accurate picture of how many language codes did not necessarily match between the LMT and Paramount codes. Those that did match were never audited again. Those that did not match were noted and sent to the LMT board.

It is important to note that just because the LMT and Paramount language codes did not always match, this did not mean either party was mistaken. The codes were simply different, and only the nomenclature agreement of the language codes would need to be normalised. For example, the LMT includes 'French as spoken in Canada', 'English as spoken in South Africa' or 'Spanish as spoken in Latin America', while the respective Paramount descriptors are 'French (Canada)', 'English (South Africa)' and 'Spanish (Latin America)'. The extent to which having a proven standard like the LMT aided when deciding which nomenclature to lean into cannot be overstated.

Finally, after confirming which language codes matched, and which did not, it was identified that Paramount was using more than 100 language codes that were not yet available on the LMT. New language codes are added to the LMT only when there is a use case available. Given that adding a new language code is always welcomed and encouraged if a use case for that specific language can be confirmed, the LMT group will review these language codes for accuracy, and once approved, they will be added to the official LMT.

Once a new LMT version is published, the intention is to bring the revised metadata language table and codes back to the stakeholders at Paramount. Communicating and establishing alignment across the organisation will establish an internal standard that will set forth a path towards consistent language codes across network enterprise systems.

Paramount's continued engagement and participation with the LMT working group will allow the inclusion of any new additions or changes to the LMT that may be put forth by other media companies. As internal technologies tend to change over time, it is important to stay close to the operations and technology leadership to ensure any changes do not affect the implementation of LMT in workflows.

LEARNING BEST PRACTICES

Organisations without a dedicated data governance team are advised to tap into the folks in the archives, library services or data science teams. These are the organisation's metadata experts and the people best positioned to start this work — and they will recognise what the LMT brings to the table. It is also important to engage with subject matter experts for the respective fields of data being reviewed.

For language, finding groups within the organisation who are dedicated to this work, or who are native speakers, is also extraordinarily helpful. Key stakeholders for this work can include research departments, technology leads (as they will have to implement tools to help do the work) and product leads. Creating a task force including these individuals can ensure clear communication and alignment companywide.

CONCLUSION

The LMT has only begun filling a void that has long existed in the media and entertainment space when it comes to the application of consistent language codes for content.

By including the same language codes all along the media supply chain, ensuring the right languages are described and delivered, consistently, all the way to when the content is delivered to viewers, the LMT serves as the Rosetta Stone for the media and entertainment industries.

Content can be stored in a media asset management and/or digital asset management system, with the language codes searched upon, and the content discovered. The LMT can also be applied to industries beyond the media and entertainment space. And as long as there is a consistent set of codes, the LMT makes it possible to better understand one another, avoid the cost and embarrassment associated with delivering the wrong language version, and eliminate the industry's Tower of Babel.

References

1. World Wide Web Consortium, (2014), 'Internet engineering task force best current practice 47', available at: https://www.w3.org/International/articles/language-tags/ (accessed 31st March, 2022).
2. MESA, (2022), 'LMT landing page', available at: https://www.mesaonline.org/language-metadata-table
3. Marshall, A., (2019), '*Mamá* to *Madre*? "Roma" subtitles in Spain anger Alfonso Cuarón', *New York Times*, 11th January, available at: https://www.nytimes.com/2019/01/11/movies/roma-spanish-subtitles-alfonso-cuaron-netflix.html (accessed 31st March, 2022).
4. IETF BCP-47, (2009), 'Tags for identifying languages', available at: https://tools.ietf.org/search/bcp47 (accessed 31st March, 2022).
5. Internet Assigned Numbers Authority, (2022), 'RFC 5646 Language sbtag registry', available at: https://www.iana.org/assignments/language-subtag-registry/language-subtag-registry (accessed 31st March, 2022).
6. SMPTE, (2015), 'Standards operations manual, version 3.1', available at: https://www.smpte.org/2020-standard-policies-and-governance (accessed 31st March, 2022).
7. IETF, (2017), 'RFC 8259 The JavaScript Object Notation (JSON) data interchange format', available at: https://datatracker.ietf.org/doc/html/rfc8259 (accessed 31st March, 2022).