Metadata is the Answer – and the Problem

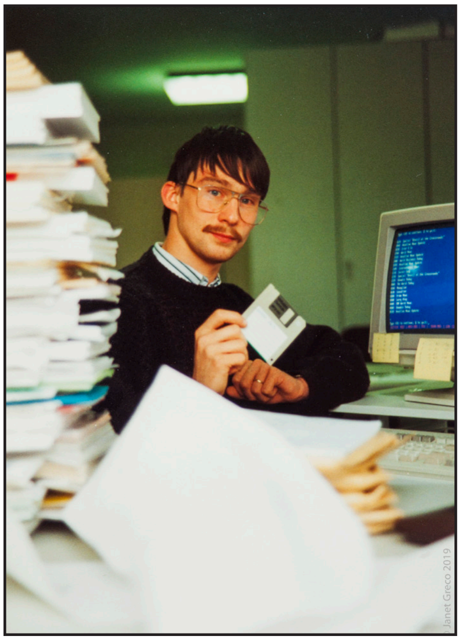
By: Janet Greco

New dynamic user interfaces and recommendations can be the difference between winning and losing in the OTT space, but managing the backbone of consumer-focused search and recommendation is a big challenge for incumbent broadcasters and content providers. The opportunity for operators will be huge once they get their metadata management practices right.

Data Driven TV. Sounds great, doesn´t it?

While OTT streaming services enjoy interrogating their data lakes with fancy dashboards, accurate audience metrics and extensive analytic capabilities, traditional broadcasters and pay TV platforms are having a tougher time of it. They face competition from the likes of Netflix, Facebook, Fortnite and everything else that vies for the attention of consumers these days.

“The role of data for operators is really very important,” says Jacques-Edouard Guillemot, SVP at Nagra, speaking at IBC 2019, who set out the issues succinctly. “They have built their entire infrastructures and business processes around data. Yet when we look at operators, they have legacy systems that for some of them are 30 years old. All departments should be able to use data in a meaningful way. But the biggest issue for our customers is that their organization is very siloed. How do you break those silos? And how do you make data available for everyone that needs it?”



Jean-François Crémer, Director, Clients Service & Operations, EMEA at TiVo, working with the incoming TV schedules from European broadcasters at Infomedia S.A. in Luxembourg, c. 1992.

While it is true that “data-driven” is the mantra of the day, the beating heart of the TV business revolves around proper metadata management in a way that “fixes” the multiple data silos within legacy operators.

Legacy innovations still in play

If the transition from paper to electronic marked the initial disruptive force in TV metadata delivery practices, the new “AI”-driven techniques for automating metadata extraction are set to be the second wave.

Artificial Intelligence (AI) is the term we often use when we really mean Machine Learning (ML) and a host of other technologies that are still at the early stages of development and deployment. These new technologies are driving a new wave of disruption in content navigation and search, enabling metadata to be automatically extracted. ML enables computer systems to learn based upon ongoing data that is provided. After a while the learning becomes more refined. AI, on the other hand is when a computer appears to learn, think and solve problems on its own. Both generate metadata *but also require* metadata to work. Beyond the AI/ML buzz, new metadata extraction technologies also include computer vision, speech to text, automatic translation, etc, and all are on-course to bring transformative benefits to the TV business.

The need for accurate TV content metadata to drive program discovery has been increasing as the number of channels and services has steadily grown. But audiences’ easy access to program information has actually decreased as we have gained more TV viewing options. As with the printed “grid” formats that preceded them, digital Electronic Program Guides (EPGs) introduced in the mid 90s made it a bit harder to access program information, because you had to click into the listing to view the description.

Janet Greco is an independent consultant with expertise in enterprise TV metadata management helping TV businesses take a more holistic view of their data operations and helping them solve problems around disparate workflows and data sets. She founded the first disruptive pan-European TV metadata aggregation service, Infomedia S.A. in 1991. Her company: www.broadcastprojects.com

Driving all these guides was the business-to-business exchange of TV metadata consisting of time, title and a program description. The first disruption in TV listings distribution took place in the early 90s with the advent of “electronic delivery” services. These intermediaries, which came to be known as TV metadata aggregators, provided a modern conduit for channels to deliver their schedules to the wide number of publications, later websites and EPGs, in an efficient manner.

What broadcasters needed at that time was someone who could sort out their TV schedule (TV metadata) distribution problems so that consumers could discover their programs. This became necessary as TV was deregulated and trans-frontier satellite broadcasting began. At that time the delivery of this information was done mainly by snail mail and fax. Programming and schedule data began to flow freely from broadcasters to aggregators. In this way, TV metadata aggregators became the necessary cog in the wheel to get linear TV data to publishers of magazines and guides, because of all the multiple formats (paper, faxes, spreadsheets, Word and text documents).

From paper to electronic, there has hardly been any further innovation for almost 30 years. Today, descriptive TV metadata is still acquired in multiple formats, aggregated and normalized. It is then delivered to anyone who needs to populate an EPG guide, including the now/next Service Information feeds (DVB-SI and PSIP in the US) which channel providers were originally meant to generate on their own.

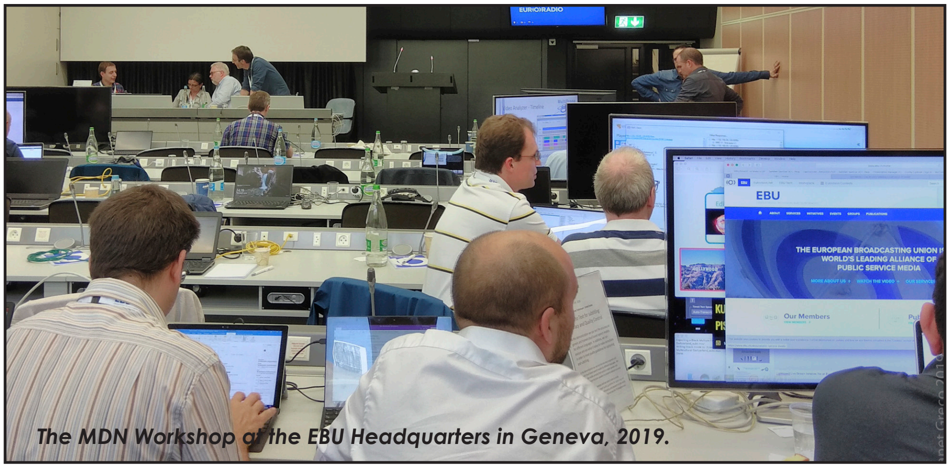
AI and Machine Learning: New Forces in TV Metadata Extraction

We are now in the midst of a second wave of disruption. A good example of this disruption can be seen at the annual workshop for developers working on metadata and artificial intelligence, hosted by the European Broadcasting Union (EBU), Europe’s representative body for public service media. The topic of media information management is part of the EBU’s strategic program on production. It aims to help members enhance and enrich their media by integrating their data (descriptive and technical but also subtitles) from commissioning to distribution.

The EBU Metadata Developer Network focuses specifically on how AI is being used to automatically extract metadata from video. It is a diverse community of academic and industry specialists working in AI, and machine learning experts from public and commercial broadcasters as well as start-ups. It provides a showcase for a variety of projects and operates under a highly collaborative atmosphere, sharing expertise and best practices.

AI can solve specific problems, like using facial recognition to speed up editing processes, enabling new forms of creativity in production, or automating the creation of subtitles and translations. Another use case is sentiment analysis, a technology already in deployment that can be very complementary for optimization and personalization of the user experience. Ultimately it means pointing the technology at the entire libraries of video content. But by what mechanism will that data make its way back to the indexed content asset?

"AI will not save the day for those who never managed data properly. Being able to attach AI generated metadata to an asset is key to everything. This is where everything begins and ends. This is particularly true and easier to manage in a semantic data framework including bringing together data from silos", says Jean-Pierre Evain, Principal Project Manager at the EBU, in charge of the EBU Strategic Programme on Media Information Management and AI.


The MDN Workshop at the EBU Headquarters in Geneva, 2019.

Data Lakes and Data Puddles

The quality of descriptive TV program information sits at the heart of a TV business. It is fundamental to the user experience, content discovery and recommendation. It is also central to AI and ML. When you have accurate data, you also stand the best chance of leveraging it. Interpreting that data based on context, categories and more granular characteristics has the best chance of success when it is meaningful and well-managed *in the first place*.

Across the industry, TV metadata sits in fragmented silos in legacy systems with no easy way to de-duplicate the data or fix the disjointed workflows that often result in time-consuming manual interventions. You could call them data puddles instead of data lakes. As many dedicated systems have been used to manage different aspects of the TV operation, the use of multiple "unique IDs" within the same organization came into play as well. These legacy systems are no joke to sort out.

So what concrete actions can be taken?

First, prioritization. Face facts and commit to getting metadata systems working in harmony to arrive at a single "meta-truth" across the enterprise. With their messy data puddles across the organization, legacy players won´t be able to keep up. In the scheme of things, SVOD and the new D2C services will be the least of their problems.

Second, be objective. Conduct a thorough review across the enterprise and assess metadata requirements per department. Bringing an objective outsider in to conduct interviews with internal stakeholders is an excellent way to break down barriers to new systems and strategic ways that support long-term business goals.

Third, approach AI technologies with meticulous attention to the devil in the details. One thing that stands out in all the demos I've seen is the fact that these systems must be meticulously trained by people! Make sure there's a clear plan so that any extracted metadata actually ends up back with the content asset, catalogued under one unique and persistent identifier. One excellent use case to consider is archive optimization, using ML, that can identify duplicate assets, assisting humans in cleaning up content libraries and the associated data more quickly.

Fourth, simplify. Use technology that simply gets the job done. Merging disparate sets of metadata can be accomplished with ordinary data analysis and management tools capable of surfacing duplicates and centralizing the multiple "unique IDs" that exist in different silos. Data can then be viewed in one single location in order to make decisions about how to edit, store, and manage this data (including editorial rules) and, where necessary, re-populate the original legacy systems that must be retained.

There is no quick fix for metadata management practices neglected in the past. During those earlier times when content owners and broadcasters began handing over their data to metadata aggregators, we arrived at a situation where aggregators now hold the keys to the kingdom, with their extensive and coherent data sets.

Ironically, today these data sets are licensed back to the companies who in turn need them to drive their guide and recommendation services. But here's the thing… those companies need data - their *own* data - for their own productions and the content for which they've acquired the rights. This same data must seamlessly tie into a host of other operations across the enterprise (ad targeting, CRM, analytics, etc). It would be a mistake to again entrust third parties to sort out metadata issues when in reality the housekeeping must be done internally first.

"With AI comes a lot more metadata that the business needs. The problem is how to store, access and use that extra data to keep the customer happy." said Willem Andries Nel, Technical Delivery Manager at Multichoice in South Africa. "A single ID would be helpful, but the underlying problem is that legacy systems will not go away overnight. We need a single data store where all the existing data and IDs are linked. This single storage would then contain all the extra AI data that the business knows we need. And then there is the problem of how to use metadata to entice Gen Z customers."

New technologies will perform best when they have the best foundational data regarding the TV content, so that people can ultimately find it. Automated metadata extraction technologies will certainly be a help, once the foundation is laid, but they are not the total solution. The consumer has an almighty problem of choice, and those choices are not just about TV anymore. If it is too difficult to find content that is relevant and relatable, chances are they can more easily find something else to switch their attention to.

It's not a question of managing churn, but being able to pre-empt it, so the only way to do that is generate engagement with your content and propose content in an intelligent and effective way. The trick is to find breakout tools that can help get this clean-up job done. Let's finally clean up those sloppy data puddles and move on. □