Artificial intelligence for a role change in television archives: The Atresmedia–Etiqmedia experience

Received (in revised form): 10th May, 2021

Eugenio López de Quintana
Head of Archive, Atresmedia Corporación de Medios de Comunicación, Spain

Eugenio López de Quintana is Head of Archive at Atresmedia, where he has been actively involved in the definition, development and implementation of in-company archive management systems and the overall processes required for the network's transition to digital. He is a former lecturer at the University of Carlos III, President of the Spanish National Association for Information and Documentation, and member of the FIAT/IFTA Executive Board.

Atresmedia Corporación de Medios de Comunicación, Avda. Isla Graciosa 13, 28703 S.S. de los Reyes, Madrid, Spain
Tel: +34 609111354; E-mail: elopez@atresmedia.com

Antonio León Carpio
Managing Director and Founder, Etiqmedia, Spain

Antonio León Carpio is the Managing Director and founder of Etiqmedia — a company specialised in the application of image, audio and text-processing technologies in archiving workflows. He has helped multiple broadcasters leverage artificial intelligence to improve their efficiency.

Etiqmedia, Calle Balbino Orensanz 55, locales 17 y 18, 50014, Zaragoza, Spain
Tel: +34 620581672; E-mail: aleon@etiqmedia.com

Abstract Given the increasing volume of audiovisual content being produced, combined with the growing demand for granularity in image searches, it is becoming increasingly impractical for television archives to catalogue their collections without automatic processing and artificial intelligence technologies. This paper describes a pioneering project to implement such technologies.

KEYWORDS: automation, information processing, artificial intelligence, audiovisual content, television archives, media content, speech recognition, facial recognition

INTRODUCTION

This paper describes a project developed by the Atresmedia Communication Group between 2019 and 2020 for the automatic processing of information through artificial intelligence solutions provided by Etiqmedia — a company specialised in the application of image, audio and text-processing technologies in archiving workflows

The Atresmedia Group comprises various free-to-air television channels, including Antena 3, LaSexta, Neox, Nova, Mega and AtresSeries, the International Channel and the online service ATRESplayer. These channels are complemented by three radio stations, Onda Cero, Europa FM and Melodía FM, and the fiction factory, Atresmedia Studios, along with Atresmedia Publicidad and Atresmedia Foundation.

It is thus little surprise that the Group is a massive generator of news programming, documentary and dramatic content, all of which must be assessed for archival purposes, along with materials of external origin, to ensure information can flow from the archive to the different lines of production.

Responsibility for managing these inward and outbound flows of content falls to the Group's documentation centre. Every year, the centre receives 60,000 hours of video content for which it must select and assign appropriate metadata to denote ownership and aid potential recovery. To this are added 65,000 hours of broadcasted programmes, of which 7,000 will be processed with a deep level of metadata to guarantee their recovery.

The image archive currently houses more than 2 million artefacts, representing a total of 3 million hours of video, with the heritage archive growing by some 14,000 hours every year. On top of this, there is also a collection of more than 400,000 photographs.

Every year, half a million video clips are transferred from the archive to different parts of the group and to production companies in different parts of the country.

THE CHANGING ROLE OF THE TELEVISION ARCHIVIST

These figures are testament to the constant increase that the audiovisual content production and consumption industry has experienced in recent years. In the more specific field of information and television, this increase has also been accompanied by a greater demand for precision and depth in database searches. This is most obvious with political information, which accounts for a substantial part of the content held by Atresmedia, and where the need to recover the exact textual phrases of different protagonists within the news scene has grown exponentially.

At the same time, however, the increase in the volume of recording hours received in the archives and the accompanying need for more precise metadata has not been accompanied by a proportional increase in the human resources required for their processing. For this reason, it is essential to look for technological alternatives that replace, as far as possible, part of the indexing and documentary description work that has traditionally been carried out manually in television archives.

The project described in this paper is the result of several years of searching for a solution, at a time when technology has begun to offer genuinely useful results.

With this project, Atresmedia has also started on the path to transform the professional profile of those working in television archives.

This will involve progressively phasing out manual work for those tasks it is possible to automate using artificial intelligence, such as recording content descriptions and people's statements. At present, more than 50 per cent of the human resources in the area are dedicated to these tasks. The hope is to redirect these resources towards new activities with the potential to add greater value, such as generating fresh content from archive materials, creating virtual universes of knowledge through ontologies that allow navigation in queries, and anticipating the information needs of users based on the news. In short, it is hoped to move archivists from processing content to generating content and information.

ALGORITHMS FOR AUTOMATIC CONTENT PROCESSING

Before describing Atresmedia's workflows, the following sections provide a brief overview of the audio, image and text processing technologies used.

Automatic speech recognition

Of all the technology described in this paper, automatic speech recognition (ASR) is probably the most commonly used.

ASR converts all voice information present in the audio into text, obtaining a list of words at the exact same time as they appear in the audio. This algorithm makes it possible to find the exact point within a piece of audio content where one can find the word, or set of words, relating to one's query.

Automatic text punctuation

This algorithm punctuates the text generated by the ASR. As text without punctuation loses readability, this algorithm is vital for helping the archivist to read and understand the ASR output. This punctuation also facilitates the subsequent segmentation of the content based on sentences and paragraphs within the text.

Speaker segmentation

This algorithm divides an audio according to the interventions of each speaker. Simply put, it marks the start and end time of each speaker intervention. This is particularly useful when segmenting content in which the interventions have semantic relevance. For example, in a press conference it makes it possible to separate the journalist's questions from the protagonist's response. It is also useful when working with interviews or radio shows.

Content segmentation

This algorithm divides long interventions into shorter segments. To do this, it works with the text of each intervention and creates clusters of sentences with the same topic. In this way, when conducting a query in the archive, it is possible recover a segment of short duration rather than working with the complete intervention.

Voice activity detection and signal-to-noise ratio in voice segments

Audio quality is a fundamental parameter for obtaining good results with the previously discussed algorithms. In the first place, the quality of recording and audio coding must be guaranteed, preferably working with sampling rates of 16 KHz and audio compressions of at least 96 Kbps. But these measures are not enough; it is also necessary to include a measure of the quality of the content itself. This entails detecting the segments of the audio with voice information and measuring their quality. If the signal-to-noise ratio is too low, the content should be discarded and tagged as invalid for automation, as it would yield poor results, leading ultimately to an archive full of noisy metadata.

Automatic subtitle resynchronisation

The subtitles generated during live broadcasts are especially useful when cataloguing such programmes. Most of these subtitles, however, have variable broadcast delays, as they are generated live and therefore cannot be perfectly synchronised. This algorithm takes an out-of-sync subtitle as input and perfectly matches the timing with the audio. This result makes it possible to find a word in the audio.

Face detection and recognition

This algorithm detects any face present in the image and stores vectors relating to characteristics, size, location and screen time. In addition, it compares this face with those stored in a previously trained repository of people, so that it can confirm whether or not the face is already in the Atresmedia database. In this way, it is possible to search the archive to retrieve content where a specific person appears in the image. One can also search for a person using a photograph of their face, and comparing vectors of characteristics with those stored in the database. With this workflow, it is possible to search for a person, even when their face has not previously been labelled with text.

NEW WORKFLOWS FOR VIDEO PROCESSING

Since the early 1990s, Atresmedia has used a document management system known as GAMA, which it developed in-house, originally to handle tape-based television production workflows. Over time, GAMA has been adapted in line with evolving digital workflows to become a powerful management system that can be integrated into any media asset management (MAM) environment, including the one currently used for the Group's various media workflows, which was also developed in-house.

The Atresmedia MAM system manages 3 million actions per month and about 60,000 video streams per week — a significant proportion corresponding to the entry and exit of archived materials.

Over the years, various automated processes have been integrated into GAMA in order to reduce manual operations in archive media workflows and the administration of their associated metadata. At the same time, automated integration processes have been developed for use with different digital production systems, as is currently the case with AVID.

Within Atresmedia's technical ecosystem, AVID is the platform for editing and generating programmes — mainly, but not limited to, news. Atresmedia's recordings, the signals from third parties and broadcasted programmes are all ingested into AVID. It is therefore one of the Group's main sources of archival content.

GAMA, which is integrated into the MAM suite, contains everything related to document management and archive workflows, from cataloguing and metadata management to all search and retrieval functionality. The MAM system integrates both the applications and the devices of the entire digital production system, including some automations that are parameterised according to needs.

The integrations and automations between AVID and GAMA are diverse, including, among other things, the import of metadata from AVID to GAMA, and direct archiving from AVID for certain types of material that may or may not have passed through the artificial intelligence algorithms. Other examples include the automatic sharing of metadata between tagged broadcast materials, clean feed copy and mastered cameras, and the automatic cleaning of pre and post images in live shows prior to archiving.

With this route of automation exhausted, the next phase in the process of reducing manual work could only come from the incorporation of automatic information analysis and artificial intelligence technologies, which took place throughout 2019 and 2020, and which will be described in due course.

One of the main requirements for the implementation of such technology in Atresmedia's archive workflows was for it to integrate into existing applications and information systems. This was essential to avoid the complexity associated with maintaining separate environments for new and legacy applications and workflows

Similarly, given the emerging nature of these new technologies and the experimental manner in which they were being developed, the scope of the project was planned to be progressive rather than comprehensive. It was also decided to use limited typologies of documents that, due to their acoustic conditions and semantic content, would be more likely to have a more favourable response to automatic processing.

In what follows, this paper will describe the four developments and workflows that are currently in operation, having been fully integrated into the archive's daily activities and management systems, with demonstrable benefits.

Complete raw materials

Raw material recordings comprise those materials recorded as part of Atresmedia's

daily coverage of current affairs, which may be combined with archival materials to produce informative programmes.

The work carried out in the archive with these materials consists of a first selection action on the recorded material, virtual generation of sublevels or sequences by semantic content, and description of the content in natural language complemented with various additional metadata (Figure 1).

The initial materials chosen for automatic processing have largely consisted of political content with either a single speaker or people speaking in turn. The content is then sent to Etiqmedia for processing, as discussed below.

The recorded raw materials are ingested in AVID, from which material to be transferred to the archive is manually selected. The selected files go through an automatic quality-control process in which metadata are also extracted to be associated with the media once the assets have been incorporated into the MAM system.

Once at the MAM system, the actions for the automatic processing and archiving of the material begin. First, a 1.5 Mb MPEG-4 H264 proxy file is generated and sent to the Etiqmedia server hosted in Atresmedia, from which a JSON file is received with all the metadata extracted by the algorithms assigned to this workflow.

The JSON file is processed by the MAM system to obtain the metadata of interest in this process and can be integrated into the GAMA data model. For each asset, GAMA incorporates a sequence structure with independent semantic content and the result of the speech-to-text transcription contained in the JSON file. From here, the archivist working with this material can review the transcript, make modifications and choose another grouping of sequences before ending the treatment.

When searching for these materials in GAMA, the content of the automatic transcription appears in a different colour from the manual descriptions, so that the user can see clearly that any minor typographical or punctuation errors that might be generated are not the result of human error. This nuance is important because it reinforces the idea that although there is no such thing as total precision in automatic information processing, it is still a very useful tool.

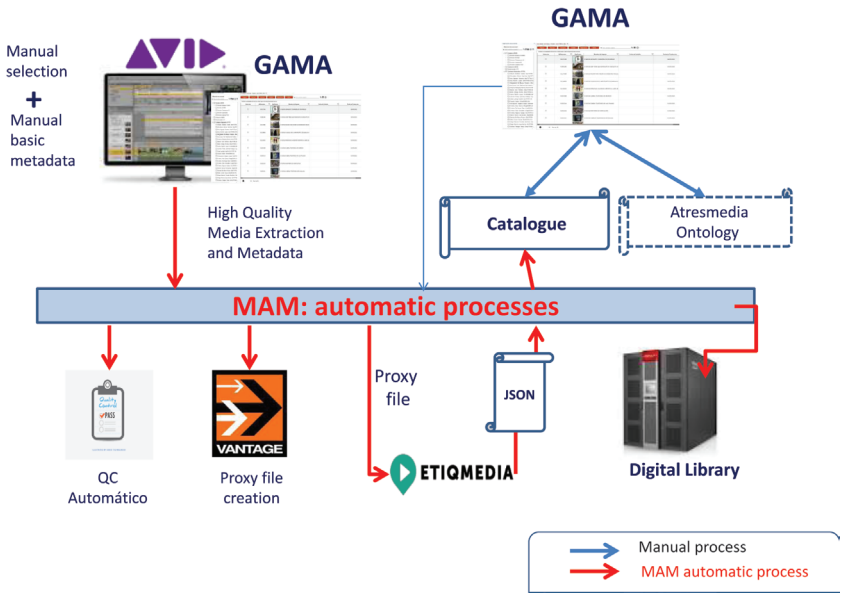

Figure 1: Automatic raw materials processing in Atresmedia's archive

The operation ends with the archiving of the proxy file on the download servers, as well as the HDCAM HD422 50 Mb file with corresponding double security redundancy according to the Atresmedia archive policy.

It is important to highlight the integration of the metadata obtained by applying algorithms in the working interfaces of the archive systems. In this way, it is possible to exercise discretion when using the algorithms. In addition, by monitoring the degree of precision obtained with each type of material, it is possible to achieve gradual progress towards automation.

Partial content of a programme

This workflow is similar to the previous one in terms of process, but it starts directly from GAMA rather than AVID. It is intended for broadcasted programmes that can only be partially processed when there is an appearance of one speaker or consecutive speakers. This is the case with interviews included in documentary and news programmes.

It is important to note, however, that algorithms are not applied with equal precision in all interviews. In these cases, both the journalistic style of the interview and the interaction between participants are decisive, as a greater tendency to constant interruptions and annotations can impair transcription and speaker segmentation.

Complete programmes by using subtitle information

The third workflow implemented is dedicated to programmes that can be processed and archived without human intervention. This represents a milestone in the archival of television programmes.

In this case, the process originates from AVID once the programme has been broadcast, and the MAM system associates the subtitle file with the proxy before sending it to the Etiqmedia server. The JSON file obtained also includes the fragmentation of sequences to be imported into GAMA, but in this case with subtitled content. For this, the Etiqmedia algorithm corrects the delay in the subtitling with respect to the time code of the video, and groups the subtitling lines into longer content units according to the requirements of the GAMA data model. As in previous cases, the description of all sequences is displayed in a different colour in the search.

Before generating the proxy, the MAM system automatically cleans the images that come before and after the live programmes, so that the final archiving of the file corresponding to the broadcasted content is guaranteed, and images that must not be reused are not included.

For these three workflows, a tracking and error correction module has been designed in GAMA which, using icons and colour codes, allows archivists to monitor the processes in progress and carry out certain forwarding actions.

Photographs through biometric patterns generation

The approach to automatic photo cataloguing is radically different. It is designed for the self-cataloguing of all photographs that are uploaded to the archive from the different areas of the Group, such as Communication, Fiction, External Production, Corporate Image, among others.

The process entails the generation of a virtual catalogue of people in GAMA in which basic metadata are assigned to each person manually, along with their position in the Atresmedia ontology and about five or six representative photographs that may come from any source (Figure 2).

These photographs are sent to the Etiqmedia server, where a biometric pattern is obtained for each of them and automatically associated with the person in the catalogue.

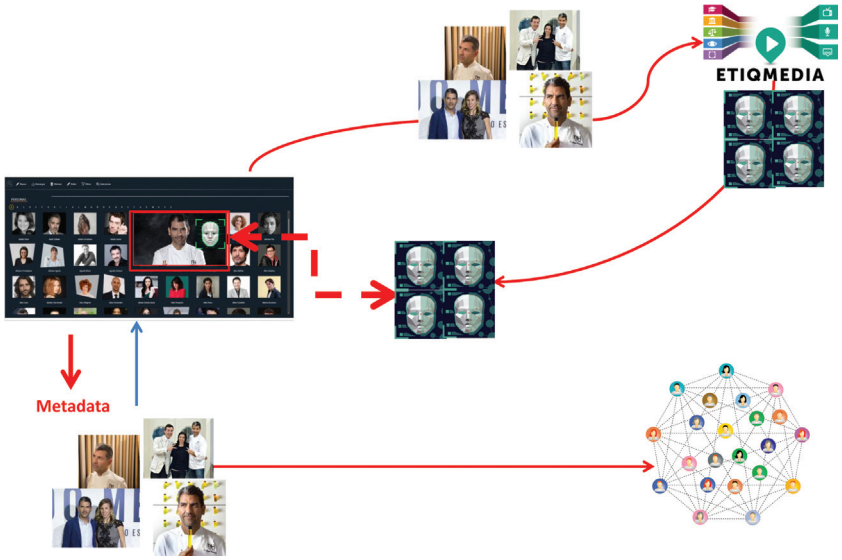

Figure 2: Workflow for cataloguing stills through the generation of biometrics patterns

In the cataloguing process, any new photograph sent to GAMA automatically inherits relevant origin metadata and is sent to the Etiqmedia server to obtain the biometric patterns of the people who appear in the photograph. In addition to the biometric pattern, the JSON file obtained also includes the position coordinates of each person in the photograph.

When GAMA processes this information, it executes an algorithmic approximation operation between the biometric patterns of each new photograph and those stored in the catalogue of people. In this way, all people represented in each photograph are not only identified but also inherit the metadata from the catalogue.

The GAMA interface offers a number of monitoring options. For example, it is possible to detect whether the pattern generation has been completed successfully and whether any people in the catalogue still need to have a pattern generated.

Within the GAMA module for photographs, it is possible to identify the person or persons that appear in a photograph simply by positioning the mouse pointer on the image.

In addition, thanks to the speed at which GAMA is able to generate patterns, the interface also supports iconic search. It is therefore possible to use a photograph from any source to build the query in the GAMA search interface. In this process, the module generates a biometric pattern or patterns for the photograph used in the query and this information is compared with the patterns housed in the file. This technology also allows Boolean combinations in the query.

TECHNOLOGY REQUIREMENTS

These workflows have made it possible to establish the technical thresholds that the system must meet before it is of actual benefit to the archivist. In this regard, it bears repeating that the objective of the algorithms is to facilitate the work of the documentalist. Thus, however interesting a result may be from a technological perspective, if it does not achieve this objective, it is not useful.

As a main requirement, it is critical to establish a specific analysis profile for each type of content. It is not feasible to have algorithms (and certainly not to train algorithms) that can work with all types of

content. For this reason, it is necessary to segment the types of content the system will process (eg political content, news content), and to define the algorithms that will be used in each use case.

When it comes to training algorithms, there is no one-size-fits-all process — each algorithm requires dedicated training. Such training significantly increases the success rates for the respective algorithms, providing results that are genuinely helpful for the archivist.

To better understand the process for adapting algorithms to use cases, the paper will show in more detail the use case implemented for political content — that is, television material relating to or generated in the political sphere, such as press conferences, rallies, statements and interviews.

The first step is to define which algorithms will be used for this type of content. With political content, the bulk of relevant information is found in the audio, as most video tracks tend to focus simply on the speaker's face. For this reason, the key algorithms relate to speech recognition, punctuation, segmentation and audio quality, while image-processing algorithms, such as object detection and scene understanding, are not particularly useful.

Once the algorithms have been selected, initial training is conducted to adapt them to the political environment. This is done by tagging previous content of the same type and training the algorithms with this tagged content. If one takes the ASR as an example, the data for training will comprise political audio content with the associated transcriptions, and large amounts of political texts, even if they do not have associated audio.

The transcribed audio is used to adapt the model from an acoustic point of view. This is especially relevant for *ad hoc* statements given in corridors or on the street, which are more likely to suffer from background noise. A good adaptation to the acoustic environment requires at least 300 hours of transcribed audio content. The text will then be used to train the language model, so it can learn the vocabulary and grammar of political content. This normally requires at least 5 million lines of text.

In the same way that training must be adapted, certain types of content require the algorithm itself to be adapted before it becomes useful for the archivist. During implementation, it is essential to respond to user feedback on an ongoing basis to ensure the system meets user needs. Using such information, it is possible to detect improvements that may be minor from a technological point of view, but which can have a significant impact on productivity. For example, Atresmedia has found that for speaker segmentation to be useful, it must have a precision level of less than one second. Likewise, automatic punctuation is critical for the content segmenter to generate relevant text clusters for the archivist.

Thanks also to user feedback, Atresmedia was able to identify that the system should not process audio content with low perceptual quality, as audio-coding information is not sufficiently powerful to make appropriate decisions. For example, content with a lot of reverberation in the audio, or with too much background noise, results in transcripts with low precision rates. Indeed, it is far better not to generate automatic metadata than to generate metadata with lots of noise.

After all this work, the system becomes a real help for archivists. Nevertheless, the work is not done — the algorithm still requires ongoing training. This is due to the emergence of new information that may not be in the initial model. In the case of ASR, for example, new terms, or previously uncommon terms, emerge all the time, with COVID-19 and coronavirus providing a case in point. Meanwhile, in the case of facial recognition, the appearance of new public figures likewise requires images of their likeness to be incorporated in the database.

RESULTS

The results obtained in these months of activity have been very satisfactory, both qualitatively and quantitatively. More than 6,000 hours of video have been processed, with an estimated reduction in the manual processing times of 20 per cent during this first measurement phase. Up to 60 hours of content can now be processed each day. Additionally, with the automatic processing of incoming media, a level of pre-cataloguing can be obtained with an immediacy that is not feasible with manual processing, and that allows for efficient recovery once cataloguing has been completed.

Regarding photographs, by the time this work is published, a substantial part of the automatic cataloguing of the documentary collection housed in GAMA will have been completed, including approximately 400,000 images that otherwise would have been unfeasible to catalogue.

The technology, however, is not without its limitations. Most obviously, the technology obtains optimal results only when content is highly structured. For example, for speech recognition algorithms to be successful, the speakers are required not to talk at the same time. They must also speak with a neutral tone. In live sports broadcasts and heated debates, where turns are not respected, the results are therefore noisy. For this reason, the main sources for the system are — at least for now — news and political content.

A further limitation relates to the audio and image quality of the input content. The quality of both the encoding and the content itself must be controlled to ensure that those files that do not meet the basic thresholds are not processed. At Atresmedia, the following minimum technical characteristics have been established: 1280×576 video resolution, H264 codec with 1.5 Mbps bitrate and 4: 2: 0 colour sampling, 16 KHz minimum audio sampling frequency and 96 Kbps audio bitrate. Regarding voice signal to noise ratio, a minimum quality threshold of 15 dB has been established.

Of course, even when the above-mentioned quality and content criteria are met, algorithms are never 100 per cent accurate, and it is necessary to manage any errors resulting from the automation. In this regard, it is essential for the algorithms to demonstrate remarkably high success rates to ensure that the combination of automated process and error management is still faster than working manually. At Atresmedia, the success threshold is set at 90 per cent. Any algorithm that cannot meet this threshold is discarded. Table 1 shows the success rates obtained using the main algorithms for each type of content.

CHANGING ROLES IN TELEVISION ARCHIVES

At present, more than 65 per cent of the human resources within the Atresmedia archive focus on the selection, cataloguing and archiving of content. With so much staff time still dedicated to little more than the transcription of political statements, a solution that combines the use of algorithms, such as those already mentioned, a consistent and up-to-date ontology, and a good search engine based on natural language, would

Table 1: Success rates of the main algorithms for each type of content

Algorithm	Type of content	Success rate (%)
ASR	Political	97.4
ASR	News	94.2
Speaker segmentation	Political	88.7
Speaker segmentation	News	86.5
Face recognition	Photographic	93.7

seem an excellent way to free up television archivists to evolve from the generation of metadata towards the management and generation of content.

A fundamental element of this transformation is to assume a certain level of heterodoxy in comparison with the strict parameters of accuracy and precision that have characterised the work of television archives to date. In return, new functions make their way, and they do so by adding to the already essential role that archives play in the content production process.

At Atresmedia, the four main lines of transformation being carried out in this regard are: (1) the creation and maintenance of ontologies to offer end users search options based on visual navigation through knowledge graphs; (2) increasing the role of archivists in programme writing teams as well as in the management of materials from external sources; (3) the generation of video content directly from archival materials; and (4) media manager functions within production systems for everything related to content workflows with or without a final destination in the archive.

NEXT STEPS

Following the success of the implemented workflows, Atresmedia is now looking to the next challenges on the horizon. For a start, there remains much to do to improve the recovery of assets based on image content. This kind of search is quite common when archival images are required to accompany a story. To date, archivists have had to label each image with a description of the scene, such as 'person with a mask on a bar terrace' and 'intensive care unit'. Scene identification algorithms would help streamline this process.

This need is currently being addressed from two angles. The first of these involves training an algorithm to label an image from a set of reference labels, thereby making it possible to conduct textual searches using the set of labels available in the thesaurus. The second approach is similar to searching for photos by face biometrics. A vector of visual characteristics of each image will be calculated, making it possible to search for similar images. That is, the search will be carried out with an image as input, and the most similar images in the archive will be retrieved.

Another requirement is to reduce the difference between the information generated by an archivist and that generated through automated processes when describing the information held in audio content. While an archivist summarises what is said in the audio and includes context, the automated process simply transcribes what is said in the audio. From an archival perspective, transcription is not the best way to file information, so Atresmedia is developing an automated system to generate abstractive summaries that provide information closer to what is needed. The creation of algorithms capable of abstractive summarisation is a complex technological challenge, and achieving this would mean a radical change in the way that archives work.

CONCLUSION

This paper has described the work carried out by Atresmedia and Etiqmedia to facilitate the documentation process within the Atresmedia Group. Experience has shown that it is critical to approach each workflow separately, selecting those types of content that can be automated rather than looking for a single solution that responds to the archive's every last need. It is also important to identify the types of content most suitable for automated processing, with a focus on content that is highly structured.

The coexistence of manual and automated workflows must be addressed. There will always be errors with automated workflows, so documentation departments must be able to manage these. For this reason, documentation departments will need to

find the right balance between manual and semi-automated flows.

The process of implementing an automation system does not end with its initial installation and setup. By its very nature, news content will always contain new information in the form of new terms, faces and types of image. To understand this information, algorithms must be trained on an ongoing basis.

As for the future of automation, the creation of algorithms capable of generating information closer to that created by archivists continues to be the goal, and the next step forward in semi-automated workflows is likely to be in the area of abstractive summarisation algorithms.